

Computational statistics using the Bayesian Inference Engine

Martin D. Weinberg¹

¹ *Department of Astronomy, University of Massachusetts, Amherst MA 01003-9305*

ABSTRACT

This paper introduces the Bayesian Inference Engine (BIE), a general parallel-optimised software package for parameter inference and model selection. This package is motivated by the analysis needs of modern astronomical surveys and the need to organise and reuse expensive derived data. I describe key concepts that illustrate the power of Bayesian inference to address these needs and outline the computational challenge. The techniques presented are based on experience gained in modelling star-counts and stellar populations, analysing the morphology of galaxy images, and performing Bayesian investigations of semi-analytic models of galaxy formation. These inference problems require advanced Markov chain Monte Carlo (MCMC) algorithms that expedite sampling, mixing, and the analysis of the Bayesian posterior distribution. The BIE was designed to be a collaborative platform for applying Bayesian methodology to astronomy. By providing a variety of statistical algorithms for all phases of the inference problem, a user may explore a variety of approaches with a single model implementation. Indeed, each of the separate scientific investigations above has benefited from the solutions posed for the other investigations, and I anticipate that the same solutions will be of general value for other areas of astronomical research. Finally, to protect one’s computational investment against loss any equipment failure and human error, the BIE includes a comprehensive persistence system that enables byte-level checkpointing and restoration. Additional technical details and download details are available from <http://www.astro.umass.edu/bie>. The BIE is distributed under the GNU GPL.

Key words: methods: data analysis - methods: numerical - methods: statistical - astronomical data bases: miscellaneous - virtual observatory tools

1 INTRODUCTION

Inference is fundamental to the scientific process. We may broadly identify two categories of inference problems: 1) *estimation*—finding the parameter of a theory or model from data; and 2) *hypothesis testing*—determining which theory, indeed if any, is supported by the data. Astronomers increasingly rely on numerical data analysis, but most cannot take full advantage of the power afforded by present-day computational statistics for attacking the inference problem owing to a lack of tools. This is especially critical when data comes from multiple instruments and surveys. The different data characteristics of each survey include varied selection effects and inhomogeneous error models. Moreover, the information content of large survey databases can in principle determine models with many parameters but exhaustive exploration of parameter space is often not feasible.

These classes of estimation problems are readily posed by Bayesian inference, which determines model parameters, θ , while allowing for straightforward incorporation of heterogeneous selection biases. In the Bayesian paradigm, current knowledge about the model parameters is expressed as a probability distribution called the *prior distribution*, $\pi(\theta)$. This is the anticipated distribution of parameters for the postulated model *before* obtaining any measurements. This should include one’s understanding of the model pa-

rameters in their theoretical context. When new data \mathbf{D} becomes available, the information content is expressed as $P(\mathbf{D}|\theta)$, the distribution of the observed data given the model parameters. This will be familiar to some as the classical *likelihood* function, $L(\mathbf{D}|\theta)$. This information is then combined with the prior to produce an updated probability distribution called the *posterior distribution*, $P(\theta|\mathbf{D})$. Bayes’ Theorem defines this update mathematically:

$$P(\theta|\mathbf{D}) = \frac{\pi(\theta)P(\mathbf{D}|\theta)}{\int \pi(\theta)P(\mathbf{D}|\theta) d\theta}. \quad (1)$$

Many will recognise this as the multiplicative rule for conditional probability. Combined with the concept of sample spaces, measure theory, and Monte Carlo computation, Bayes theorem provides a rich framework for the quantitative investigation of a wide variety of inference problems, such as classification and cluster analyses, which broadly extends the two groups described above. Later sections will illustrate the importance and utility of explicit quantification of the prior information.

Why use the Bayesian framework? To begin, the Bayesian approach unifies both aspects of the inference problem: estimation and hypothesis testing. For example, given a galaxy image and several families of brightness profiles, we would like to determine both the distribution of parameters for each family and which family is best supported by the data. A classical analysis would report the

arXiv:1203.3816v1 [astro-ph.IM] 16 Mar 2012

Maximum Likelihood (ML) estimate for each model using a χ^2 -type statistic (Pearson 1900) and prefer the fit with the lowest value of χ^2 per degree of freedom. However, the χ^2 statistic will grow with sample size in the presence of measurement errors. This leads to the well-known over-fitting problem where the Pearson-type χ^2 test will reject the correct distribution in favour of one which better describes the deviations caused by the measurement errors. This well-known issue may be treated in a variety of ways, but the Bayesian approach naturally *prefers* the model with the smallest number of dimensions that can explain the data distribution through the specification of the prior information. The Bayesian approach further emphasises that model comparison problems must depend on the prior distribution. See §2.2 for more details.

The computational complexity of applying equation (1) directly grows quickly with the number of model parameters and becomes intractable before the volume of currently available large data sets is reached. However, Monte Carlo algorithms based on Markov chains for drawing samples from the posterior distribution promise to make the Bayesian approach very widely applicable (e.g. see Robert & Casella 2004). In turn, the application of Bayesian methods in cosmology and astrophysics has flourished over the past decade, spurred by data sets of increasing size and complexity (Trotta 2008). Once a scientist can determine the posterior distribution, rigorous credible bounds on parameters and powerful probability-based methods for selecting between competing models and hypotheses immediately follow. This statistical approach is superior to those commonly used in astronomy because it makes more efficient use of all the available information and allows one to test astronomical hypotheses directly. To realise this promise for astronomical applications, we need a software system designed to handle both large data sets and large model spaces simultaneously.

Beginning in 2000, a multidisciplinary investigator team from the Departments of Astronomy and Computer Science at UMass designed and implemented the Bayesian Inference Engine¹, a Markov chain Monte Carlo (MCMC) parallel software platform for performing statistical inference over very large data sets. We focused on probability-based Bayesian statistical methods because they provide maximum flexibility in incorporating and using all available information in a model-data comparison. For example, multiple data sources can be naturally combined and their selection effects, which must be specified by the data provider to obtain a meaningful statistical inference, are easily incorporated. In this way, the BIE provides a platform for investigating inference using the virtual observatory paradigm.

This paper has multiple goals. I begin by introducing in §2 the concepts in Bayesian inference that illustrate its power as a framework for parameter estimation and model selection for astronomical problems. This power, not surprisingly, comes with significant computational challenges that informs our design for the BIE; the features of the package are described in Appendix A. I then describe and motivate our choice of MCMC algorithms (§3); the BIE allows additional algorithms to be straightforwardly added as needed. This is followed by a brief summary of BIE-enabled research in §4. I summarise in §5.

¹ See <http://www.astro.umass.edu/bie> for detailed description and download instructions.

2 WHAT DO ASTRONOMERS WANT AND NEED?

2.1 Parameter estimation

Many astronomical data analysis problems are posed as parameter estimates. For example: 1) one measures a spectral energy distribution of an object and would like estimate its temperature; or 2) one measures the flux profile of a disk galaxy and would like to estimate its scale length. In these problems, one is asserting that the underlying model is true and testing the hypothesis that the parameter, temperature or scale length, has a particular value.

Bayesian inference approaches these problems with the following three steps, reflecting the standard practice of the scientific method: 1) numerically quantify a prior belief in the hypothesis; 2) collect data that will either be consistent or inconsistent with the hypothesis; 3) compute the new confidence in the hypothesis given the new data. These steps may be repeated to achieve the desired degree of confidence. A clever observer will design campaigns that refine confidence efficiently (i.e., that makes the confidence high or low). In the context of our simple examples, one may believe that the spectral energy distribution is that of an M-dwarf star and one's prior belief is then a distribution of values centred on 2000 K. After measuring the spectral energy distribution, the prior distribution of temperature is combined with the probability of observing the data for a particular temperature, to get a refined distribution of the temperature of the object. Notice that this procedure does not result in a single value. Rather, the posterior probability distribution is used to estimate a credible interval (also known as a Bayesian confidence interval). Of course, credible intervals and regions are only a simple summary of the information contained in the posterior distribution. Unlike classical statistics, Bayesian inference does not rely on a significance evaluation based on theoretical or empirical reference distributions that are valid in the limit of very large data sets. Rather it specifies the probability distribution function for the parameters explicitly based on the data at hand.

A prime motivation for the BIE project is the thesis that the power of expensive and large survey data sets is underutilised by targeting parameter estimation as the goal. To illustrate this, let us consider the second example above: estimating the scale length of a disk. A standard astronomical analysis might proceed as follows. One determines the posterior probability distribution for scale lengths for some subset of survey images. Alongside scale length, one determines other parameters such as luminosity, axis ratios, or inclinations, and possibly higher moments such as the asymmetry. The scale length with maximum probability becomes the *best estimate* and is subsequently correlated with some other parameter of interest, luminosity or asymmetry, say. Then, any correlation is interpreted in the context of theories of galaxy formation and evolution. Observe, that in the first step, one is throwing out much of the information implicit in the posterior distribution. In particular, the luminosity estimate is most likely correlated with the scale-length estimate. If one were to plot the posterior distribution in these two parameters, one might find that the distribution is elongated in the scale-length–asymmetry plane, possibly in the same sense as the putative correlation! In other words, the confidence in the hypothesis of a correlation should include the full posterior distribution of parameter estimates, not just the maximum probability estimate. See §4.2 and Figure 3 for a real-world example.

Moreover, this scenario suggests that one is using disk scale length and asymmetry as a proxy for testing a hypothesis about disk evolution or environment. These results might have been more reliable if the observational campaign had been designed to enable a

hypothesis test, not a parameter estimate, from the beginning. This leads naturally to the following question.

2.2 Which model or theory is correct?

This question is a critical one for the scientific method. Astronomers typically do not address it quantitatively but *want* to do so. I will separate the general question “which model is correct?” into two: 1) “does the model explain the data?”, the *goodness-of-fit* problem; and 2) “which of two (or more) models better explains the data?”, the *model selection* problem. Let us begin with (1) and discuss (2) in the next section.

Suppose one has performed a parameter estimation and determined the parameter region(s) containing a large fraction of the probability. Before making any conclusions from the application of a statistical model to a data set, an investigator should assess the fit of the model to make sure that the model can explain adequately the important aspects of the data set. *Model checking*, or assessing the fit of a model, is a crucial part of any statistical analysis. Serious misfit (failure of the model to explain important aspects of the data that are of practical interest) should result in the replacement or extension of the model. Even if a model has been assumed to be final, it is important to assess its fit to be aware of its limitations before making any inferences.

The posterior predictive check (PPC) is a commonly-used Bayesian model evaluation method (e.g. Gelman et al. 1995, Chap. 6). It is simple and has a clear theoretical basis. To apply the method, one first defines a set of discrepancy measures. A discrepancy measure, like a classical test statistic, measures the difference between an aspect of the observed data set and the theoretically predicted data set. Let \mathcal{M} denote the model under consideration. Practically, a number of predicted data sets are generated from $P(\mathbf{D}|\theta^*, \mathcal{M})$ with θ^* selected from the posterior distribution. Any systematic differences between the observed data set and the predicted data sets indicate a potential failure of the model to explain the data. For example, one may use the distribution of a discrepancy measure based on synthetic data generated from the posterior distribution to estimate a Bayesian p-value for the true data under the model hypothesis. The p-value in this context is simply the cumulative probability for the discrepancy statistic. A p-value in the tails of the predicted discrepancy-measure distribution suggests a poor fit to the data. By using a variety of different discrepancy statistics, one’s understanding of *how* the model does not fit the data is improved. See §3.5.1 for more detail.

Another approach attempts to fit a non-parametric model to the data. If the non-parametric model better explains the data than the fiducial model, one rejects the fiducial model as a good fit. A procedure for assessing the model families will be described in the next section. A naive implementation of this idea is difficult, requiring a second high-dimensional MCMC simulation to infer the posterior distribution for the non-parametric model and a careful specification of the prior distribution. A clever scheme for doing this (Verdinelli & Wasserman 1998) is described in §3.5.2.

2.3 Model selection and Bayes factors

One often has doubts about our parametric models, even those that fit. This is especially true when the models are phenomenological rather than the results of *first-principle* theories. Therefore, one needs to estimate which competing model better represents the data. This approach has been applied to good advantage in interpreting the spatial fluctuations in the cosmic microwave back-

ground radiation. The Bayesian model selection approach has been shown to efficiently answer questions such as: what is the best fit combinations of baryonic, dark matter and dark energy components? See Trotta (2008) for a review.

Astronomers are becoming better versed in the more traditional statistical *rejection* tests but astronomers often really want *acceptance* tests. Bayes factors provide this: one can straightforwardly evaluate the evidence *in favour* of the null hypothesis rather than only test evidence for rejecting it. *Bayes factors* are the dominant method for Bayesian model selection and are analogous to likelihood ratio tests (e.g. Jeffreys 1961; Gelman et al. 1995; Kass & Raftery 1995). Rather than using the posterior extremum, one marginalises over the parameter space to get the marginal probability of the data under each model or hypothesis. The ratio of the likelihood functions marginalised over the prior distributions provides evidence in favour of one model specification over another. In this way, the Bayesian approach naturally includes, requires in fact, that one’s prior knowledge of the model and its uncertainties be included in the inference. Although this dependence on the prior probability sometimes criticised as a flaw in the Bayesian approach, I believe this is as it should be: one’s prior belief will invariably influence one’s interpretation of a statistical finding. The Bayesian framework allows the scientist to describe and incorporate prior beliefs quantitatively. In addition, the method demands that the scientist thoughtfully characterise prior assumptions to start; such discipline will improve the quality of any scientific conclusions.

Bayes factors are very flexible, allowing multiple hypotheses to be compared simultaneously or sequentially. The method selects between models based on the evidence from data without the need for nesting². The posterior probability for competing models can be evaluated over an ensemble of data and used to decide whether or not a particular family of models should be preferred. Similarly, common parameters can be evaluated over a field of competing models with appropriate posterior model probabilities assigned to each. A tutorial illustrating this can be found in the BIE documentation.

Mathematically, Bayes factors follow from applying Bayes Theorem to a space of models or hypotheses. Let $P(\mathcal{M})$ be our prior belief in Model \mathcal{M} , and let $P(\mathbf{D}|\mathcal{M})$ be the probability of observing \mathbf{D} under the assumption of Model \mathcal{M} . The Bayes Theorem tells us that the probability of Model \mathcal{M} having the observed \mathbf{D} is

$$P(\mathcal{M}|\mathbf{D}) = \frac{P(\mathcal{M})P(\mathbf{D}|\mathcal{M})}{P(\mathbf{D})} \quad (2)$$

where $P(\mathbf{D})$ is some unknown normalisation constant. However, one may use equation (2) to compute the relative probability of two competing models, \mathcal{M}_i and \mathcal{M}_j :

$$\frac{P(\mathcal{M}_1|\mathbf{D})}{P(\mathcal{M}_2|\mathbf{D})} = \frac{P(\mathcal{M}_1) P(\mathbf{D}|\mathcal{M}_1)}{P(\mathcal{M}_2) P(\mathbf{D}|\mathcal{M}_2)} \quad (3)$$

without reference to the unknown normalisation. The left-hand side of equation (3) may be interpreted as the posterior odds ratio of Model 1 to Model 2. Similarly, the first term on the right-hand side is the prior odds ratio. The second term on the right-hand side is called the Bayes factor. Most often, one does not assert a preference for either model and assigns unity to the prior odds ratio.

To define Bayes factors explicitly in terms of the posterior distribution, suppose that one observes data \mathbf{D} ; these may comprise

² Two models are *nested* if they share the same parameters and one of them has at least one additional parameter.

Table 1. Jeffreys' table

$\log B_{12}$	B_{12}	Strength of evidence
< 0	< 1	Negative (supports M_2)
0 to 1/2	1 to 3.2	Barely worth mentioning
1/2 to 1	3.2 to 10	Positive
1 to 2	10 to 100	Strong
> 2	> 100	Very strong

many observations or multiple sets of observations. One wishes to test two competing models (or hypotheses) M_1 and M_2 , each described by its own set of parameters, θ_1 and θ_2 . One would like to know which of the following likelihood specifications is better: $M_1 : L_1(\mathbf{D}|\theta_1)$ or $M_2 : L_2(\mathbf{D}|\theta_2)$, given the prior distributions $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ for θ_1 and θ_2 . The Bayes Factor B_{12} is given by

$$B_{12} = \frac{P(\mathbf{D}|\theta_1)}{P(\mathbf{D}|\theta_2)} = \frac{\int \pi_1(\theta_1|M_1)P_1(\mathbf{D}|\theta_1, M_1)d\theta_1}{\int \pi_2(\theta_2|M_2)P_2(\mathbf{D}|\theta_2, M_2)d\theta_2}. \quad (4)$$

If $B_{12} > 1$, the data indicate that M_1 is more likely than M_2 and vice versa. Harold Jeffreys (1961, App. B) suggested the often-used scale for interpretation of B_{12} in half-unit steps in $\log B_{12}$ (see Table 1). This provides a simple-to-use, easily discussed criterion for the interpretation of Bayes factors. Note that classical hypothesis testing gives one hypothesis (or model) preferred status (the *null hypothesis*) and only considers evidence against it; the Bayes factor approach is considerably more general.

Given all of these advantages, why are Bayes factors not more commonly used? There are two main difficulties. First, multidimensional integrals are difficult to compute. Following equation (4), one needs to evaluate an integral of the form: $P(\mathbf{D}) = \int \pi(\theta)P(\mathbf{D}|\theta)d\theta$. For a real world model, the dimensionality of θ is likely to be > 10 . Such a quadrature is infeasible using standard techniques. On the other hand, a typical MCMC calculation has generated a large number of evaluations of the integrand at considerable expense. Can one use the posterior sample to evaluate the integral?

Raftery (1995) suggests a *Laplace-Metropolis* estimator that uses the MCMC posterior simulation to approximate the marginal density of the data using Laplace's approximation (see Raftery op. cit. for details). In practice, this is only accurate for nearly Gaussian (or *normal*, eq. B1) unimodal posterior distributions. As part of the BIE development, Weinberg (2012) described two new approaches for evaluating the marginal likelihood from the MCMC-generated posterior sample and both of these are implemented in the BIE as secondary analysis routines (see §3.2). In short, I believe that the BIE together with recent advances for computing the marginal likelihood makes the wholesale computation of Bayes factors feasible in many cases of interest.

A second well-known difficulty is the sensitivity of Bayes factors to the choice of prior. Most commonly, researchers feel that vague priors are more appropriate than informative priors. This leads to an inconsistency known as the Jeffreys–Lindley paradox (Lindley 1957), which shows that vague priors result in overwhelming odds for or against a hypothesis by varying the parameter that controls the vagueness (e.g. extending the range of an arbitrary uniform distribution). This apparent problem has led researchers to seek Bayesian hypothesis tests that are less sensitive to their prior distributions. Conversely, I argue that one should not expect a vague prior to yield a sensible model comparison. Rather, the prior should be used to express prior belief in a theory and, therefore, the resulting hypothesis test should be sensitive to the prior. This sensitivity

implies that the theory implicit in the model is informed by one's background knowledge. Nonetheless, the prior knowledge is difficult to quantify and I would still advocate testing a variety of prior distributions consistent with one's prior knowledge. This may be tested through direct sensitivity analyses, such as resimulation with chains at different resolutions and approximate priors.

Regardless of one's viewpoint, I believe that the BIE project currently provides a useful platform for investigating the use of Bayesian model comparison and hypothesis testing and hope that it will help pave the way for new applications. In some cases, computing the Bayes factor will be infeasible. For these, the BIE includes an MCMC algorithm that selects between models as part of the posterior simulation (reversible jump) as described in §3.3.

2.4 Observational requirements

The probability of the data given the parameter vector and the model, $P(\mathbf{D}|\theta, M)$ or the likelihood function, is fundamental to any inference, Bayesian or otherwise. Meaningful inferences demand that the data presentation include all of the information necessary for the modeller to compute $P(\mathbf{D}|\theta, M)$ accurately and precisely. The more direct the construction of $P(\mathbf{D}|\theta, M)$ from the physical theory, i.e. the less information lost in modeling the acquired observations, the easier it is to calculate $P(\mathbf{D}|\theta, M)$, leading to a quality result. In other words, the more the data is reduced through summary statistics and “cleaned” by applying complex filters, the less information remains and the greater the effect of difficult-to-model correlations. This is somewhat contrary to standard practice.

In addition, astronomers often quote their error models in the form of uncorrelated standard errors. The customary expectation is that each datum, typically a data bin or pixel, should be within the range specified by the error bar most of the time. Quoted error bars are often inflated to make this condition obtain. This leads to a number of fundamental flaws that makes the error model (and therefore the data) unsuitable for Bayesian inference:

(i) Binned and pixelated data are nearly always correlated. For example, a flat-field photometric correction correlates the pixels of an image over its entire scale. Sky brightness removal has similar effects. There are many additional sources of indirect correlations. Parameter estimations are often sensitive to these correlated excursions in the data values and ignoring these correlations will lead to erroneous inferences. Data archivists can facilitate accurate inferences by providing correlation matrices for all error models.

(ii) Selection effects must be modeled in the likelihood function and, therefore, these effects must be well specified by the archivist to facilitate straightforward computation. For example, consider a multiband flux-limited source catalogue. A colour-magnitude or Hess diagram in two flux bands will have a non-rectangular boundary owing to the flux limit. Although this is a simple example, selection effects may be terribly difficult to model; consider spatial variations in source completeness owing to the diffraction spikes from bright stars.

(iii) Astronomers tend to use historically familiar summary data representations that inadvertently complicate the computation of $P(\mathbf{D}|\theta, M)$. Continuing the previous example, the magnitude-magnitude diagram contains the same information as the colour-magnitude diagram but the selection effects lie along flux-level boundaries. For a more complicated example, consider the Tully-Fisher diagram. The input data set may contain flux limits, mor-

phology selections, image inclination cuts, redshift range limits, just to name a few.

In summary, data processing and reduction, correlates the data representation and complicates the computation of $P(\mathbf{D}|\boldsymbol{\theta}, \mathcal{M})$. This renders the modeling process difficult and potentially unreliable. Rather, one should endeavour to separate each source of error, carefully specifying the underlying acquisition process for every observational campaign. For example, each pixel datum in a digital image may be characterised by the observed data number (dn), gain and bias, read noise, thermal background fraction, etc. Together, their combination yields an error model. Even if a first-principle process cannot be described, an empirical distribution or process description will be helpful to modelers. For example, a distribution of deviations for measured pixel values relative to a reference calibration field may be used as an error model. Ideally, the data representation should be as close to the acquired form as possible. In cases where the archiving or presentation of source data is impractical, the production of a correlation matrix is essential.

The effect of data correlation has been explored by Lu et al. (2011). They describe the parameter inference for a semi-analytic model of galaxy formation conditioned on a galaxy mass function with both correlated and uncorrelated data bins. The differences in the posterior distributions for these two cases is dramatic. When the error model is in doubt, the sensitivity of the inference to the error model can be investigated in the Bayesian paradigm by putting prior distributions on parameters of the error models that describe their uncertainty, marginalising over those hyperparameters, and comparing with the original posterior. Although this is more expensive and rarely done, one should consider performing such sensitivity analyses regularly.

3 SOLUTIONS PROVIDED BY THE BIE

Assume that one has identified the observational selection effects, specified the prior distributions, and constructed $P(\mathbf{D}|\boldsymbol{\theta}, \mathcal{M})$. How does one compute and the posterior distribution defined by equation (1) using MCMC?

3.1 Computing the posterior distribution

This section presents four MCMC sampling algorithms of increasing complexity included in the BIE. I begin with a description and some motivation for the standard Metropolis-Hastings algorithm. This simple easy-to-use and implement algorithm often fails to converge and is difficult to tune for complicated high-dimension distributions. The next three sections introduce modifications that circumvent these pitfalls.

3.1.1 The Metropolis-Hastings algorithm

Metropolis-Hastings is the most well-known of MCMC algorithms (Metropolis et al. 1953; Hastings 1970). This algorithm constructs a Markov chain that generates states from the $P(\boldsymbol{\theta}|\mathbf{D}, \mathcal{M})$ distribution after a sufficient number of iterations. Its success requires that the Markov chain satisfy both an ergodicity and a detailed balance condition. The ergodicity condition ensures that at most one asymptotic distribution exists. Ergodicity would fail, for example, if the chain could cycle back to its original state after a finite number iterations. Ergodicity also requires that all states with positive probability be visited infinitely often in infinite time (called *positive*

recurrence). For general continuous state spaces, ergodicity is readily achieved. The detailed balance condition ensures that the chain admits at least one asymptotic distribution. Just as in kinetic theory or radiative transfer, one defines a transition process that takes an initial state to a final state with some probability. If the Markov chain samples the desired target distribution $P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{D}, \mathcal{M})$, detailed balance demands that the rate of transitions $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$ is the same as the rate of transitions from $\boldsymbol{\theta}' \rightarrow \boldsymbol{\theta}$.

One may state this algorithm explicitly as follows. Let $P(\boldsymbol{\theta})$ be the desired distribution to be sampled and $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ be a known, easy-to-compute transition probability between two states. Given $\boldsymbol{\theta}$, the distribution $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is a probability distribution for $\boldsymbol{\theta}'$. Let $a(\boldsymbol{\theta}, \boldsymbol{\theta}')$ be the probability of accepting state $\boldsymbol{\theta}'$ given the current state $\boldsymbol{\theta}$. In short, one can show that if the detailed balance condition

$$P(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\theta}')a(\boldsymbol{\theta}, \boldsymbol{\theta}') = P(\boldsymbol{\theta}')q(\boldsymbol{\theta}', \boldsymbol{\theta})a(\boldsymbol{\theta}', \boldsymbol{\theta}) \quad (5)$$

holds, then the Markov chain will sample $P(\boldsymbol{\theta})$. It is straightforward to verify by substitution that acceptance probability

$$a(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min \{1, [P(\boldsymbol{\theta}')q(\boldsymbol{\theta}', \boldsymbol{\theta})/P(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\theta}')]\} \quad (6)$$

solves this equation (for additional discussion see Liu 2004). Equation (5) has the same form as well-known kinetic rate equations as follows. Given the probability of a transition over some time interval from a state A to some other state B of a physical system and the corresponding reverse reaction, then the equilibrium condition for $N = N_A + N_B$ systems distributed in the two states is:

$$N_A P(A \rightarrow B) = N_B P(B \rightarrow A).$$

In equation (5), the probability densities $P(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}')$ play the rôle of the occupation numbers N_A and N_B and the product of the transition and acceptance probabilities play the rôle of the probabilities $P(A \rightarrow B)$ and $P(B \rightarrow A)$.

The transition probability $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is often chosen to facilitate the generation of $\boldsymbol{\theta}'$ from $\boldsymbol{\theta}$. Metropolis et al. (1953) introduced a kernel-like transition probability $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \bar{q}(\boldsymbol{\theta} - \boldsymbol{\theta}')$ where $\bar{q}(\cdot)$ is a density. This has the easy-to-use property of generating $\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \bar{q}$. (Here and throughout, the notation $\boldsymbol{\theta} \sim P(\boldsymbol{\theta})$ expresses that the distribution function of the variate $\boldsymbol{\theta}$ is $P(\boldsymbol{\theta})$.) Further, if \bar{q} is symmetric, i.e. $\bar{q}(\mathbf{z}) = \bar{q}(-\mathbf{z})$, then equation (6) takes the simple form

$$a(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min \left\{ 1, \frac{P(\boldsymbol{\theta}')}{P(\boldsymbol{\theta})} \right\} \quad (7)$$

The BIE provides two symmetric distributions for \bar{q} : a multivariate normal and a uniform or *top-hat* distribution. Each element of $\boldsymbol{\theta}$ is scaled by a supplied vector of *widths*, \mathbf{w} . The choice of \mathbf{w} is critical to the performance of the algorithm. If the width elements are too large, $P(\boldsymbol{\theta}')/P(\boldsymbol{\theta})$ will tend to be very small and proposed states will rarely be accepted. Conversely, if the width elements are too small, the new state frequently will be accepted and successive states will be strongly correlated. Either extreme leads to process that is slow to reach equilibrium. The optimal choice is somewhere in between the two. As the dimensionality of the parameter space grows, specifying the optimal vector of widths a priori is quite difficult. I will address this difficulty in §3.1.4.

3.1.2 Tempered transitions

In addition to the inherent difficulties associated with *tuning* the transition probability, the Metropolis-Hastings state can easily be trapped in isolated modes, between which the Markov chain moves

only rarely. This prevents the system from achieving detailed balance and, thereby, prevents sampling from the desired target distribution $P(\theta)$. There are a number of techniques for mitigating this so-called *mixing problem*. For the BIE, I adopted a synthesis of Metropolis-coupled Markov chains (Geyer 1991) and a simulated tempering method proposed by Neal (1996) called *tempered transitions*. To sample from a distribution $P_0(\theta) \equiv P(\theta)$ with isolated modes, one defines a series of n other distributions, $P_1(\theta), \dots, P_n(\theta)$, with P_k being easier to sample than P_{k-1} . For example, one may choose

$$P_k(\theta) \propto P_0^{\beta_k}(\theta) \quad (8)$$

with $1 = \beta_0 > \beta_1 > \dots > \beta_{n-1} > \beta_n > 0$. This construction has a natural thermodynamic interpretation. One may write $P_0^{\beta_k}(\theta) = e^{\beta_k \log(P_0)} \equiv e^{\log(P_0)/T_k}$. The distribution with temperature $T_0 = T = 1$ is the original distribution. *Hotter* distributions have higher temperature $T_k > T_0$ and are over-dispersed compared with the original *cold* distribution. In other words, taking a distribution function to a small fractional power decreases the dynamic range of its extrema. In the limit $T_k \rightarrow \infty$, P_k becomes uniform. Next, the method defines a pair of base transitions for each k , \hat{T}_k and \check{T}_k , which both have P_k as an invariant distribution and satisfy the following mutual reversibility condition for all θ and θ' : $P_k(\theta)\hat{T}_k(\theta, \theta') = \check{T}_k(\theta', \theta)P_k(\theta')$.

A tempered transition first finds a candidate state by applying the base transitions in the sequence $\hat{T}_1 \cdots \hat{T}_n$. After each upward transition, new states are sampled from a broader distribution. In most cases, this liberates the candidate state from confinement by the mode of the initial state. This is then followed by a series of downward transitions $\check{T}_n \cdots \check{T}_1$. This candidate state is then accepted or rejected based on ratios of probabilities involving intermediate states. Thermodynamically, each level k corresponds to an equilibrium distribution at temperature T_k . Therefore, the upward or downward transitions correspond to *heating* or *cooling* the system, respectively. One chooses the maximum temperature to be sufficiently high to *melt* any structure in the original *cold* posterior distribution that would inhibit mixing.

Explicitly, the algorithm proceeds as follows. The chain begins in state $\hat{\theta}_0$ and obtains the candidate state, $\check{\theta}_0$, as follows. For $j = 0$ to $n - 1$, generate $\hat{\theta}_{j+1}$ from $\hat{\theta}_j$ using \hat{T}_{j+1} . Set $\hat{\theta}_n = \hat{\theta}_n$. Then, for $j = n$ to 1 , generate $\check{\theta}_{j-1}$ from $\check{\theta}_j$ using \check{T}_j . The candidate state, $\check{\theta}_0$, is then accepted with probability

$$a \equiv \min \left[1, \frac{P_1(\hat{\theta}_0)}{P_0(\hat{\theta}_0)} \cdots \frac{P_n(\hat{\theta}_{n-1})}{P_{n-1}(\hat{\theta}_{n-1})} \cdot \frac{P_{n-1}(\check{\theta}_{n-1})}{P_n(\check{\theta}_{n-1})} \cdots \frac{P_0(\check{\theta}_0)}{P_1(\check{\theta}_0)} \right]. \quad (9)$$

Although equation (9) looks complicated, the derivation in Neal (1996) shows that it immediately follows by recursively applying equations (5) and (6). Then, owing to the mutual reversibility condition, the \hat{T} and \check{T} dependence in equation (9) cancels. If the candidate state is not accepted, the next state of the Markov chain is the same as the original state, $\hat{\theta}_0$. In practice, I ‘burn-in’ the chain at each level j for M iterations, with $M = 20$ typically. Note that each P_i occurs an equal number of times in the numerator and denominator in equation (9). Therefore, the acceptance probability can be computed without knowledge of the normalisation constants for these distributions. If the acceptance probability is to be reasonably high, properly-spaced intermediate distributions will have to be provided that gradually interpolate from P_0 to P_n . Thermodynamically, this corresponds to an adiabatic increase of the heat-bath temperature followed by an adiabatic return to the original temperature. The maximum temperature $T_n = \beta_n^{-1}$ should be large enough to break any barriers between modes at low temperature

but not so large that $P_n(\theta)$ provides no constraint on the distribution of θ . Since this value depends on the features of $P(\theta)$ that are not known a priori, I choose T_n by checking the distribution of θ for P_n with various values of T_n . Our tests have shown that this algorithm works remarkably well for a variety of different inference problems.

Many readers will be familiar with the idea of *simulated annealing* (Kirkpatrick et al. 1983). Each step of a simulated annealing algorithm proposes to replace the current state by a state randomly constructed from the current state by a transition probability. The magnitude of the excursion allowed by the transition probability is controlled by a temperature-like parameter that is gradually decreased as the simulation proceeds. This prevents the state from being stuck at local maxima. The tempered transitions algorithm is heuristically similar to simulated annealing with the important additional property of obeying detailed balance.

3.1.3 Parallel tempering

The parallel tempering algorithm inverts the order of the previous algorithm: it simultaneously simulates n chains, each with target distribution P_j (eq. 8) and proposes to swap states between adjacent members of the sequence at predefined intervals. The high temperature chains are generally able to sample large volumes of parameter space, whereas low temperature chains, may become trapped in local probability maxima. Parallel tempering achieves good sampling by allowing the systems at different temperatures to exchange states at very different locations in parameter space. The higher temperature chains often achieve detailed balance quickly and accelerate the convergence of the lower temperature chains. Thus, this method may allow a simulation to achieve detailed balance even in the presence of widely separated modes. In some situations, parallel tempering outperforms tempered transitions with a lower overall computational cost. In addition, the parallel tempering algorithm is trivially parallelised by assigning each chain its own process. The tempered transitions algorithm is intrinsically serial and parallel efficiency is only obtained if the likelihood computation is parallelisable.

The parallel tempering algorithm proceeds as follows. At each step, a pair of adjacent simulations in the series is chosen at random and a proposal is made to swap their parameter states, whose acceptance is determined using a Metropolis-Hastings criterion. Let the j^{th} iterate of the state in the k^{th} chain be denoted as $\theta_j^{[k]}$. The swap is accepted with probability

$$a = \min \left[1, \frac{P_k(\theta_j^{[k+1]}|\mathbf{D}, \mathcal{M})P_{k+1}(\theta_j^{[k]}|\mathbf{D}, \mathcal{M})}{P_k(\theta_j^{[k]}|\mathbf{D}, \mathcal{M})P_{k+1}(\theta_j^{[k+1]}|\mathbf{D}, \mathcal{M})} \right], \quad (10)$$

where $P_k(\theta|\mathbf{D}, \mathcal{M})$ is the posterior probability of θ given the data \mathbf{D} for chain k and model assumptions \mathcal{M} . Final results are based on samples from the $\beta_0 = 1$ chain. As in the tempered transitions algorithm, the high-temperature states will mix between separated modes more efficiently, and subsequent swapping with lower-temperature chains will promote their mixing.

The analog thermodynamic system here is an *array* of systems with the same internal dynamics at different temperatures. At higher temperatures, strongly ‘forbidden’ states are likely to remain forbidden but valleys between multiple modes are likely to be more easily crossed. In contrast to tempered transitions (§3.1.2), the proposed transitions in parallel tempering are *sudden* exchanges of state between systems of possibly greatly different temperature. As with tempered transitions, the algorithm obeys the detailed balance equation.

3.1.4 Differential evolution

Real-world high-dimensional likelihood functions often have complex topologies with strong anisotropies about their maxima (see §4.1, Fig. 2). Difficulties in tuning the Metropolis-Hastings transition probability to achieve both a good acceptance rate and good mixing plagues high-dimensional MCMC simulations of the posterior probability. This problem affects all of algorithms discussed up to this point. Recently Ter Braak (2006) introduced an MCMC variant of a genetic algorithm called *differential evolution* (Price 1997; Storn & Price 1997; Storn 1999). This version of differential evolution uses an ensemble of chains, run in parallel, to adaptively compute the Metropolis-Hastings transition probability. In all that follows, I will refer to the MCMC variant as simply differential evolution.

The algorithm proceeds as follows. Assume that our ensemble has n chains to start, e.g., initialised from the prior probability distribution. Each chain has the same target distribution $P(\theta)$. The original differential evolution algorithm (Price 1997) proposes to update member i as follows: $\theta_p^{[j]} = \theta_{R0}^{[j]} + \gamma(\theta_{R1}^{[j]} - \theta_{R2}^{[j]})$ where $R0, R1, R2$ are randomly selected without replacement from the set $\{1, 2, \dots, n\}$. The proposal vector replaces the chosen one if $P(\theta_p^{[j]}) > P(\theta_i^{[j]})$. Ter Braak (2006) shows that with minor modifications the transition probability and the acceptance condition for differential evolution obeys detailed balance. The new MCMC version of the differential evolution algorithm takes the form $\theta_p^{[j]} = \theta_i^{[j]} + \gamma(\theta_{R1}^{[j]} - \theta_{R2}^{[j]}) + \epsilon$ where ϵ is drawn from a symmetric distribution with a small variance compared to that of the target, but with unbounded support such as a d dimensional normal distribution (eq. B1) with very small variance σ^2 : $\epsilon \sim \mathcal{N}^d(0, \sigma^2)$. The random variate ϵ is demanded by the recurrence condition: the domain for non-zero values of the posterior P must be reached infinitely often for an infinite length chain. The proposal $\theta_p^{[j]}$ is accepted as the next state for Chain i with probability

$$a = \min \left[1, \frac{P(\theta_p)}{P(\theta_i)} \right].$$

In essence, differential evolution uses the variance between a population of chains whose distributions are converging to the target distribution to automatically tune the proposal widths. Although the transition probability distribution $q(\theta, \theta')$ does not have an analytic form in this application, the differential evolution algorithm enforces symmetry through the random choice of indices, and the distribution $q(\theta, \theta')$ clearly exists.

This algorithm as stated above does not address the mixing problem. Ter Braak (2006) suggests including simulated tempering or simulated annealing in differential evolution. Along these lines, BIE also includes a hybridised differential evolution which periodically performs a tempered transition step for all n chains in parallel. This provides an ensemble at each temperature for the upward and downward transitions. As in tempered transitions, we evolve each chain for M steps at each temperature level. A typical number of temperature levels is twenty, and therefore, the addition of tempering may slow the algorithm by an order of magnitude. Although *tempered* differential evolution is dramatically slower than the simple differential evolution, I have found it essential for achieving a converged posterior sample for many of our real-world astronomical inference problems. This method was applied to the Bayesian semi-analytic models described in Lu et al. (2011).

3.1.5 Summary: choice of a MCMC algorithm

I advocate performing a suite of preliminary simulations to explore the features of one's posterior distribution with various algorithms. My experience suggests that there is no single *best* MCMC algorithm for all applications. Rather, each choice represents a set of trade offs: more elaborate algorithms with multiple chains, augmented spaces, etc., are more expensive to run but may be the only solution for a complex posterior distribution. Conversely, an elaborate algorithm would be wasteful for simulating a simple posterior distribution. Moreover, combinations of MCMC algorithms in multiple-chain schemes are often useful.

For distributions with complex topologies, differential evolution relieves the scientist of the task of hand selecting a transition probability by trial and error. This method has the advantage of added efficiency: states from all chains in a converged simulation provide valid posterior samples. However, this strategy may backfire if the posterior is strongly multimodal because differential evolution requires multiple chains in each mode to enable mixing between modes. Any single chain in a discrete mode will remain forever. Parallel chains and similar algorithms do not have this problem. Although high-temperature chains from tempered methods do not sample the posterior distribution, they do provide useful information for importance sampling. Yoon et al. (2012) has productively used high-temperature samples for Monte Carlo integration.

3.2 Computation of Bayes factors and marginal likelihoods

As described in §2.3, the marginal likelihood plays a key role in Bayesian model selection. There are several common strategies for computing the marginal likelihood. The simplest is direct quadrature using multidimensional cubature algorithms (e.g. Berntsen et al. 1991). Computational complexity limits its application to approximately four or fewer dimensions. Secondly, one may approximate the integrand around each well-separated mode as a multivariate normal distribution and integrate the resulting approximation analytically. This is the *Laplace approximation*. It suits simple unimodal densities approximately Gaussian shape, but the posterior distributions for many real-world problems are far from Gaussian. Finally, one may rewrite Bayes theorem as an expression that evaluates the normalisation constant from the posterior sample as the harmonic mean of the likelihood function, as will be shown below. In short, none of the three suffices in general: direct quadrature is most often computationally infeasible, the Laplace approximation works well only for simple posterior distributions and the harmonic mean approximation often has enormous variance owing to its inverse weighting by the likelihood value (see Kass & Raftery 1995). To help address this lack, Weinberg (2012) presents two computationally-modest families of quadrature algorithms that use the full sample posterior but without the instability of the harmonic mean approximation (Newton & Raftery 1994) or the specificity of the Laplace approximation (Lewis & Raftery 1997).

The first algorithm begins with the normalised Bayes theorem:

$$Z \times P(\theta|\mathbf{D}) = \pi(\theta)L(\mathbf{D}|\theta) \quad (11)$$

where

$$Z \equiv \int_{\Omega} d\theta P(\theta|\mathbf{D}) = \int_{\Omega} d\theta \pi(\theta)L(\mathbf{D}|\theta) \quad (12)$$

normalises $P(\theta|\mathbf{D})$ (as in eq. 1). The quantity Z is called the *normalisation constant* or *marginal likelihood* depending on the con-

text. Dividing by $L(\mathbf{D}|\boldsymbol{\theta})$ and integrating over $\boldsymbol{\theta}$ we have

$$Z \times \int_{\Omega} d\boldsymbol{\theta} \frac{P(\boldsymbol{\theta}|\mathbf{D})}{L(\mathbf{D}|\boldsymbol{\theta})} = \int_{\Omega} d\boldsymbol{\theta} \pi(\boldsymbol{\theta}). \quad (13)$$

Since the Markov-chain samples the posterior, $P(\boldsymbol{\theta}|\mathbf{D})$, the computation of the integral on the left from the chain appears as an inverse weighting with respect to the likelihood. This is poorly conditioned owing to the inevitable small values of $L(\mathbf{D}|\boldsymbol{\theta})$. However, if the integrals in equation (13) are dominated by the domain sampled by the chain, the integrals can be approximated by quadrature over a truncated domain, Ω_s that eliminates the small number of the chain states with low $L(\mathbf{D}|\boldsymbol{\theta})$. More precisely, the integral on the left-hand-side may be cast in the following form:

$$\int_{\Omega_s} d\boldsymbol{\theta} \frac{P(\boldsymbol{\theta}|\mathbf{D})}{L(\mathbf{D}|\boldsymbol{\theta})} = \int dY M(Y). \quad (14)$$

This integral will be a good approximation to the original if the measure function defined by

$$M(Y) = \int_{1/L(\mathbf{D}|\boldsymbol{\theta}) > Y} d\boldsymbol{\theta} P(\boldsymbol{\theta}|\mathbf{D}). \quad (15)$$

decreases faster than L as $L \rightarrow 0$. Otherwise, the integral in equation (14) does not exist and the first algorithm cannot be used; see Weinberg (2012) for details. Intuitively, one may interpret this construction as follows: divide up the parameter space $\boldsymbol{\theta} \in \Omega_s$ into volume elements sufficiently small that $P(\boldsymbol{\theta}|\mathbf{D})$ is approximately constant within each volume element. Then, sort these volume elements by their value of $Y(\mathbf{D}|\boldsymbol{\theta}) \equiv L^{-1}(\mathbf{D}|\boldsymbol{\theta})$. The probability element $dM \equiv M(Y + dY) - M(Y)$ is the prior probability of the volume between Y and $Y + dY$. However, if the truncated volume forms the bulk of the contribution to equation (14), the evaluation will be inaccurate.

To evaluate the r.h.s. of equation (13), one may use the sampled posterior distribution itself to tessellate the sampled volume in $\Omega_s \subset \Omega$. This may be done straightforwardly using a space-partitioning structure. A binary space partition (BSP) tree, which divides a region of parameter space into two exclusive sub regions at each node, is particularly efficient. The most easily implemented tree of this type for arbitrary dimension is the kd-tree (short for k-dimensional tree). The kd-tree algorithms split \mathbb{R}^k on planes perpendicular to one of the coordinate system axes. The implementation provided for the BIE uses the median value along one of axes (a *balanced* kd-tree). I have also implemented a hyper-octree. The hyper-octree generalises the octree by splitting each n-dimensional parent node into 2^n hypercubic children. Unlike the kd-tree, the hyper-octree does not split on point location and the size of the cells is not strictly coupled to the number of points in the sample. In addition, the cells in the kd-tree might have extreme axis ratios but the cells in the hyper-octree are hypercubic. This helps provide a better representation of the volume containing sample points. See Weinberg (2012) for additional details, tests, and discussion. Approximate tessellations also may be useful. For example, the nearest neighbour to every point in a sample could be used to circumscribe each point by a sphere of maximum volume such that all spheres are non-overlapping. Comparisons and performance details will be reported in a future contribution (Yoon et al. 2012).

For cases where the integral in (14) does not exist or the first algorithm provides is a poor approximation, Z may be evaluated directly using the second computational approach. Begin by integrating equation (11) over $\Omega_s \subset \Omega$:

$$Z \times \int_{\Omega_s} d\boldsymbol{\theta} P(\boldsymbol{\theta}|\mathbf{D}) = \int_{\Omega_s} d\boldsymbol{\theta} \pi(\boldsymbol{\theta}) L(\mathbf{D}|\boldsymbol{\theta}). \quad (16)$$

The Monte Carlo evaluation of the integral on the left-hand side is simply the fraction of sampled states in Ω_s relative to the entire sample: $F_{\Omega_s} \equiv \sum_{\boldsymbol{\theta}_i \in \Omega_s} 1 / \sum_{\boldsymbol{\theta}_i \in \Omega} 1$. The integral on the right-hand side may be evaluated using the space-partitioning procedure described above. Altogether, then, one has

$$Z = F_{\Omega_s}^{-1} \int_{\Omega_s} d\boldsymbol{\theta} \pi(\boldsymbol{\theta}) L(\mathbf{D}|\boldsymbol{\theta}) \quad (17)$$

where Ω_s is ideally chosen to avoid regions of very low posterior probability. This method has no problems of existence for proper probability densities.

There are several sources of error in this space partition. For a finite sample, the variance in the tessellated parameter-space volume will increase with increasing volume and decreasing posterior probability. This variance may be estimated by bootstrap. As usual, there is a variance-bias trade-off in choosing the resolution of the tiling: the bias of the probability value estimate increases and the variance decreases as the number of sample points per volume element increases. Some practical examples suggest that the resulting estimates are not strongly sensitive to the number of points per cell.

In summary, the choice between the various algorithms depends on the problem at hand. The Laplace approximation may be a good choice for posterior distributions that are unimodal with light tails but this is often not the case for real-world problems. I investigate the performance of the algorithms in Weinberg (2012) for high-dimensional distributions in Yoon et al. (2012). To date, I have reliably evaluated Z for $n \leq 14$ using a MCMC-generated samples of approximately 10^6 points with auto-correlation lengths of approximately 20. Additional performance details and tests in astronomical applications are described in Yoon et al. (2012).

3.3 Dimension switching algorithms

When generating large sample sizes that are necessary for an accurate computation of the marginal likelihood is impractical, one may propose a number of different models and choose between them as part of the Monte Carlo sampling process. This is done by adding a discrete indicator variable to the state to designate the active model. The resulting state space consists of a discrete range for the indicator and of continuous ranges for each of the parameters in each model. Green (1995) showed that the detailed balance equation can be formulated in such a general state space. This allows one to propose models of different dimensionality and thereby incorporate model selection into the probabilistic simulation itself. The algorithm requires a transition probability to and from each subspace (Green 1995).

For example, suppose one has an image of a galaxy field that one would like to model with some unknown number of distinct galaxies $k \in [1, n]$. The extended sample space, then, consists of n subspaces, each one of which contains the parameter vectors for each of the k galaxies for each subspace. Suppose that the current state is in the $k = 3$ subspace; that is, the current model has three galaxies. With some predefined probability at each step, $p(3, 4)$, the algorithm proposes a transition to $k = 4$ galaxies by splitting one of three galaxies into two separate but possibly blended components. Similarly, with some predefined probability at each step, $p(4, 3)$, one defines the reverse transition by combining two adjacent components into one component. Finally, no subspace transition is proposed with the probability $p(3, 3) = 1 - p(3, 4) - p(4, 3)$. For model comparison, an estimate of the marginal probability for each model k follows directly from the occupation frequency in each subspace of the extended state space.

To make this explicit following Green (1995), one first defines reversible transitions between models in different subspaces, say i and j . This is accomplished by proposing a bijective function g_{ij} that transforms the parameters between subspaces $g_{ij}(\theta^{[i]}, \phi^{[i]}) = (\theta^{[j]}, \phi^{[j]})$, and enforces the dimensional matching condition $d(\theta^{[i]}) + d(\phi^{[i]}) = d(\theta^{[j]}) + d(\phi^{[j]})$ where the operator $d(\cdot)$ returns the rank of the vector argument. The parameter vector $\phi^{[i]}$ is a random quantity used in proposing changes in the components and for choosing additional components when going to a higher dimension. The rank of $\phi^{[i]}$ may be zero. For example, if $d(\theta^{[i]}) = 2$ and $d(\theta^{[j]}) = 1$ then one may define $d(\phi^{[i]}) = 0$ and $d(\phi^{[j]}) = 1$. In other words, for the purposes of inter-dimensional transitions, each subspace is augmented by random variates with the constraint that the dimensionality of the augmented spaces match.

Then, if $q_{ij}(\theta^{[i]}, \phi^{[i]})$ is the probability density for the proposed transition and $p(i, j)$ is the probability to move from subspace i to subspace j , the acceptance probability may be written as

$$\alpha_{ij}(\theta^{[i]}, \phi^{[i]}) = \min \left\{ 1, \frac{P_j(\theta^{[j]}|\mathbf{D})p(j, i)q_{ji}(\theta^{[j]}, \phi^{[j]})}{P_i(\theta^{[i]}|\mathbf{D})p(i, j)q_{ij}(\theta^{[i]}, \phi^{[i]})} \left| \frac{\partial(\theta^{[j]}, \phi^{[j]})}{\partial(\theta^{[i]}, \phi^{[i]})} \right| \right\} \quad (18)$$

where the final term in the second argument of $\min\{\cdot\}$ is the Jacobian of the mapping between the augmented spaces, and $P_j(\theta^{[j]}|\mathbf{D})$ is the posterior probability density for the model in subspace j . The probability densities $p(i, j)$ and q_{ij} are selected based on prior knowledge of the problem and to optimise the overall rate of convergence. The algorithm can be summarised as follows. Assume that the state at iteration i is in subspace n_i with parameter vector $\theta_i^{[n_i]}$. One proposes a new state as follows:

- (i) Choose a new model j by drawing it from distribution $p(n_i, \cdot)$. Propose a value for the parameter $\theta^{[j]}$ by sampling $\phi^{[n_i]}$ from the distribution $q_{n_i j}(\theta_i^{[n_i]}, \phi^{[n_i]})$.
- (ii) Accept the move with probability $\alpha_{n_i j}(\theta_i^{[n_i]}, \theta^{[j]})$.
- (iii) If the move is accepted, let $n_{i+1} = j$ and $\theta_{i+1}^{[n_{i+1}]} = \theta^{[j]}$.
- (iv) If the move is not accepted, stay in the current subspace : $n_{i+1} = n_i$ and $\theta_{i+1}^{[n_{i+1}]} = \theta_i^{[n_i]}$.

Green (1995) named this algorithm *Reversible Jump Markov chain Monte Carlo* (RJMCMC). One often chooses to interleave RJMCMC steps with some number of standard Metropolis-Hastings steps to improve the mixing in each subspace.

Within the constraints, the transitions are only limited by one's inventiveness. However, I have found that the most successful transitions are intuitively motivated by features of the scientific problem. Although RJMCMC enables model selection between arbitrary families of models with various dimensionality, the convergence of the MCMC simulation depends on the specification of the transition probabilities and a careful tuning of the MCMC algorithm. In addition, I have not tried RJMCMC with multiple-chain algorithms that propose transitions between chains. This appears to be formally sound but the requirement that the transitioning chains be in the same subspace may yield very long mixing times. Similarly, the differential evolution algorithm (see §3.1.4) would require that each subspace of interest be populated by some number of chains all times.

Table 2. RJMCMC applied to the images in Figure 1

Model	A_3/A_2	A_3/A_{total}	$p(2)^\dagger$	$p(3)$	$p(4)$
1	1.0	0.167	0.0	0.9997	0.0003
2	0.5	0.091	0.0	0.9997	0.0003
3	0.3	0.057	0.444	0.556	0.0003
4	0.2	0.038	0.963	0.036	0.001
5	0.0	0.0	0.998	0.002	0.0

[†] $p(m)$ is the probability of states in the subspace with m components. For these simulations, $p(1) = p(5) = p(6) = 0$

3.3.1 Example: a simple transition probability

Consider two models, Model 1 with two real parameters: $\theta^{[2]} : (\theta_1, \theta_2) \in \mathcal{R}^2$ and Model 2 with one real parameter $\theta^{[1]} : \theta \in \mathcal{R}$. Let us assume that the prior probability of transition between the models is $p(1, 2) = p(2, 1) = 1/2$ and adopt the transformation between the subspaces: $g_{12}(\theta_1, \theta_2) = [(\theta_1 + \theta_2)/2, (\theta_1 - \theta_2)/2] = (\theta, \phi)$. The variable ϕ is distributed as $q(\phi)$. The inverse transformation is: $g_{21}(\theta, \phi) = (\theta + \phi, \theta - \phi)$. Therefore, given θ , one draws ϕ from $q(\phi)$ and immediately obtain (θ_1, θ_2) . The acceptance probability for $(\theta_1, \theta_2) \rightarrow \theta$ becomes

$$\alpha_{21} = \frac{\pi_1(\theta)q(\phi)}{\pi_2(\theta_1, \theta_2)} \left| \frac{\partial(\theta, \phi)}{\partial(\theta_1, \theta_2)} \right| = \frac{\pi_1(\frac{\theta_1+\theta_2}{2})q(\frac{\theta_1-\theta_2}{2})}{2\pi_2(\theta_1, \theta_2)}. \quad (19)$$

and for $\theta \rightarrow (\theta_1, \theta_2)$:

$$\alpha_{12} = \frac{2\pi_2(\theta + \phi, \theta - \phi)}{\pi_1(\theta)q(\phi)}. \quad (20)$$

3.3.2 Example: BIE mixture modeling

Many problems in astronomy are mixtures of components drawn from the same model family. Each component j in the mixture is additively combined with a weight w_j such that $\sum_{j=1}^m w_j = 1$. This allows the predefinition of some generic RJMCMC transitions that are likely to work for a wide variety of problems. I consider two types of transitions. The first is the *birth* and *death* of a component. The birth step is implemented by selecting a new component from the prior distribution for the component parameters for the user-specified model. The prior for the weights is chosen here to be a Dirichlet distribution (eq. B4) with a single user-specified shape parameter since each component is assumed to be indistinguishable a priori. Assume that the new component is born in state with m components. The new $m + 1$ weight is selected from the prior distribution for $m + 1$ weights after marginalising over m of the them. The m weights in the current state are scaled to accommodate the choice. The death of a component is the inverse of the birth step defined by detailed balance. The second type of transitions are *split* and *join* transitions; again, these are inverses. For the split step, one selects a component at random and splits each component with an additive or multiplicative shift as follows: $\theta_1 = \theta_0 + \delta\theta$, $\theta_2 = \theta_0 - \delta\theta$, or $\theta_1 = \theta_0(1 + \epsilon)$, $\theta_2 = \theta_0(1 - \epsilon)$.

For an example, consider a toy model for a group of galaxies where the light from each galaxy has a normal distribution with the same width. Each image has two well-separated components with $A_1 = 4000$ and $A_2 = 1000$ counts. A tertiary component with five different amplitudes A_3 is added along the line between the two, separated by a half width. With this choice, the secondary and tertiary components are blended and appear as a single peak (see Fig. 1 for details). The results of applying the reversible jump transitions above to these four images and one with

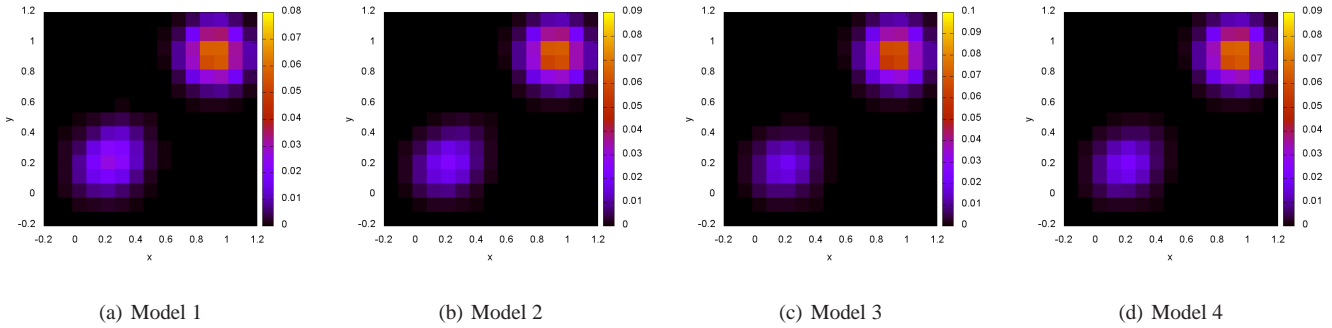


Figure 1. Four test models with three two-dimensional Gaussian distributions of width 0.1 in each dimension. The three components have centres at (0.2, 0.2), (0.9, 0.9), (0.3, 0.3). The first and second components for all models were realised with 1000 and 4000 counts respectively. The third component for Models 1, 2, 3 and 4 have 1000, 500, 300, and 200 counts, respectively. Components Two and Three are barely distinguishable by eye, even for Models 1 and 2 and appear as an elongation. The “by eye” differences between Models 3, 4 and Model 5 with zero tertiary amplitude are not easily distinguished.

an empty tertiary component is described in Table 2. The total number of counts is $A_{total} = \sum_{i=1}^3 A_i$ and $p(m)$ denotes the probability of states in subspace m . The prior probability on the centre coordinates is uniform in the range $(-0.2, 1.2)$. The prior probability on each subspace with $m \in [1, 6]$ components is Poisson with a mean of 1; that is, I am intentionally preferring fewer components, but tests with Poisson means of 2 and 3 show that the results are insensitive to this choice. The MCMC simulations use the tempered transitions algorithm described in §3.1.2 with $T_{max} = 8$ and 16 logarithmically-spaced temperature levels. The Metropolis-Hastings transition probability is a uniform ball whose widths are adjusted to achieve an acceptance rate of approximately 0.15. The RJMCMC algorithm puts nearly all of the probability on the two- and three-component subspaces. The posterior distribution of component centres are accurately constrained to the input values in the correct subspace. Table 2 reveals a sharp transition between preferring three to two components at an amplitude ratio of the tertiary to secondary amplitude, A_3/A_2 , at 0.3 (Model 3) and below. Figure 1 shows that this procedure reflects our expectation that RJMCMC should identify three separate components when one can do so by eye. Of course, the subtle visual asymmetries that allows us to do this would be obscured by noise that would be included in a realistic model.

In summary, RJMCMC allows one to identify the number of components in a mixture without using Bayes factors. The simple set of transitions used here are likely to work well for a wide variety of model families since they depend on the mixture nature, not properties of the underlying model families. Unlike Bayes factors, RJMCMC simulations may not be reused for model comparisons in light of a new model; rather, the RJMCMC simulation must be repeated including the new model.

3.4 Convergence testing

The BIE provides extensible support for convergence testing. Convergence testing has two goals: (1) determining when the simulation is sampling from the posterior distribution, and (2) determining the number of samples necessary to represent the distribution. Here, I address the first goal. For multiple chains, the work horse is the commonly used Gelman & Rubin (1992) statistic. This method compares the interchain variance to the intrachain variance for an ensemble of chains with different initial conditions; the similarity of the two is a necessary condition for convergence.

For single-chain algorithms, I have had good success with a diagnostic method that assesses the convergence of both marginal and joint posterior densities following Giakoumatos et al. (1999, hereafter GVDP). This method determines confidence regions for the posterior mean by batch sampling the chain. As the distribution converges, the distribution of the chain states about the mean will approach normality owing to the central limit theorem: the variance as a function of $1/\sqrt{N}$ for a sample size N will be linearly correlated for a converged simulation. This approach generalises Gelman & Rubin (1992) who used the coefficient of determination (C.O.D., the square of the Pearson’s product-moment correlation coefficient, e.g. Press et al. 1992, §14.5) to assess convergence. Moreover, GVDP use the squared ratio of the lengths of the empirical estimated confidence intervals for the parameters of interest as an alternative interpretation of the R diagnostic. This alternative calculation of R is simpler to compute than the original ratio of variances and is free from the assumption of normality.

For parallel chain applications, and especially for differential evolution (see §3.1.4), some chains will get stuck in regions of anomalously low posterior probability. Several outliers in a large ensemble of chains can be removed without harming the simulation as long as the chains are independent.

3.5 Goodness-of-fit testing

As described in §2, model assessment is an essential component of inference. I have explored two approaches: Bayesian p -values and Bayes factors for a non-parametric model. The first is easily applied but is a qualitative indicator only. The second has true power as a hypothesis test but is computationally intensive.

3.5.1 Posterior predictive checking

Once one has successfully simulated the posterior distribution, one may predict future data points easily. The predicted distribution of some future data \mathbf{D}^{pred} after having observed the data \mathbf{D} is

$$\begin{aligned} p(\mathbf{D}^{pred}|\mathbf{D}) &= \int p(\mathbf{D}^{pred}, \boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \\ &= \int p(\mathbf{D}^{pred}|\boldsymbol{\theta}, \mathbf{D}) p(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta}, \end{aligned} \quad (21)$$

called the *posterior predictive distribution*. In the last integral expression in equation (21), $p(\mathbf{D}^{pred}|\boldsymbol{\theta}, \mathbf{D})$ is the probability of ob-

serving \mathbf{D}^{pred} given the model parameter θ and observed data set \mathbf{D} . The distribution $p(\theta|\mathbf{D})$ is the posterior distribution. For many problems of interest, the probability of observing some new data given the model parameter θ is independent of the original \mathbf{D} . In many cases, $p(\mathbf{D}^{pred}|\theta, \mathbf{D})$ will be the standard likelihood function $p(\mathbf{D}^{pred}|\theta)$, simplifying equation (21). One may simulate the posterior predictive distribution using an existing MCMC sample as follows: 1) sample m values of θ from the posterior; 2) for each θ in a posterior set, sample a value of \mathbf{D}^{pred} from the likelihood $p(\mathbf{D}^{pred}|\theta)$. The m values of \mathbf{D}^{pred} represent samples from the posterior predictive distribution $p(\mathbf{D}^{pred}|\mathbf{D})$.

One can attempt to check specific model assumptions with *posterior predictive checks* (PPC) using the posterior predictive distribution. The idea is simple: if the model fits, predicted data generated under the model should look similar to the observed data. That is, the discrepancy measure applied to the true data should *not* lie in the tails of the predicted distribution. If one sees some discrepancy, does it owe to an inappropriate model or to random variance? To answer this question (following Gelman 2003, and references therein), one generates M data sets, $\mathbf{D}_1^{pred}, \dots, \mathbf{D}_M^{pred}$ from the posterior predictive distribution $p(\mathbf{D}^{pred}|\mathbf{D})$. Now one chooses some number of test statistics $T(\mathbf{D}, \theta)$ that measure the discrepancy between the data and the predictive simulations. These discrepancy measures can depend on the data \mathbf{D} and the parameters and hyperparameters θ , which is different from standard hypothesis testing where the test statistic only depends on the data, but not on the parameters. The discrepancy measures $T(\mathbf{D}, \theta)$ need to be chosen to investigate deviations of interest implied by the *nature* of the problem at hand. This is similar to choosing a powerful test statistic when conducting a hypothesis test. Any chosen discrepancy measure must be meaningful and pertinent to the assumption you want to test. Examples of this approach using the BIE may be found in Lu et al. (2012).

3.5.2 Non-parametric tests

This class of goodness-of-fit tests weights a parametric null hypothesis against a non-parametric alternative. For example, one may wish to test the accuracy of an algorithm that has produced n independent variates $\theta_{1:n} = (\theta_1, \theta_2, \dots, \theta_n)$ intended to be normally distributed. One would test the null hypothesis that the true density is the normal distribution $\mathcal{N}(\mu, \sigma^2)$ against a rich non-parametric class of densities by placing a prior distribution on the null and alternative hypotheses and calculating the Bayes factor. This leads to difficult, high-dimensional calculations.

For the BIE, I adopt a remarkably clever method proposed by Verdinelli & Wasserman (1998, hereafter VW) to perform a non-parametric test without proposing alternative models directly. The VW approach is based on the following observation: since the cumulative distribution function for a variable θ , $F(\theta)$, is strictly increasing and continuous, the inverse $F^{-1}(u)$ for $u \in [0, 1]$ is the unique real number θ such that $F(\theta) = u$. In the multivariate case, the inverse of the cumulative distribution function will not be unique generally, but, instead, one may define

$$F^{-1}(u) = \inf_{\theta \in \mathbb{R}^d} \{F(\theta) \geq u\} \quad (22)$$

for a parameter vector θ of rank d . Then, rather than defining a general class of densities in \mathbb{R}^d to propose the alternative, VW consider a functional perturbation to F , $G(F(\theta))$ say, such that G maps the unit interval onto itself. The identity, $G(u) = u$, is the unperturbed probability distribution. Then, the test evaluates the uniformity of

Table 3. Marginal likelihood values for Verdinelli-Wasserman tests described in §3.5.2

Class	Model type	$\ln P(\mathbf{D} \mathbf{M})$	B_{12}^\dagger
VW (1)	Gaussian	$586.7^{+1.6}_{-0.01}$	$-1.4^{+1.6}_{-0.02}$
Fiducial (2)	Gaussian	$588.1^{+0.01}_{-0.01}$	
VW (1)	Power law	$588.7^{+4.1}_{-3.5}$	$474.2^{+4.1}_{-3.5}$
Fiducial (2)	Power law	$114.5^{+0.01}_{-0.01}$	

[†] Following the definition in §2.3, B_{12} denotes the odds that Model 1 is more likely than Model 2.

the distribution of probabilities under each hypothesis. To be concrete, let F_0 describe the probability of sample θ under hypothesis H_0 . That is, F_0 is the cumulative distribution function of the posterior distribution. Under H_0 , one expects each random variate $F_0(\theta_j)$ for all j to be independent and identically distributed (iid) as a uniform distribution in the unit interval. This may be notated as follows:

$$H_0 : F_0(\theta_1), F_0(\theta_2), \dots, F_0(\theta_n) \stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1).$$

Alternatively, a poor fit will satisfy the hypothesis H_1 :

$$H_1 : F_0(\theta_1), F_0(\theta_2), \dots, F_0(\theta_n) \not\stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1).$$

To construct the functional perturbation G , Verdinelli & Wasserman use a sequence of Legendre polynomials, $\{\xi_j(\cdot), j = 1, 2, \dots\}$, defined over the unit interval to construct infinite exponential densities of the form

$$g(u|\psi) = \exp \left[\sum_{j=1}^{\infty} \psi_j \xi_j(u) - c(\psi) \right]$$

$\psi = (\psi_1, \psi_2, \dots)$ are coefficients and

$$c(\psi) = \log \int_0^1 du \exp \left[\sum_{j=1}^{\infty} \psi_j \xi_j(u) \right]$$

is a normalising constant. VW put priors on ψ , given by $\psi_j \sim \mathcal{N}(0, \tau^2/c_j^2)$ where τ and the c_j are appropriately chosen constants. VW also specify a hyperprior on τ : $\tau \sim \mathcal{N}(0, w^2)$ truncated to the positive values. This distribution provides finite probability for obtaining the null hypothesis near $\tau = 0$ and decreases monotonically for larger perturbations from the null, maintaining the perturbative nature of the alternative hypothesis.

Intuitively, this development is closely related to the probabilistic interpretation of the marginal likelihood and Bayes factors. To see this, consider the one-dimensional case for simplicity: let $f(\mathbf{D}|\theta) = P(\theta)P(\mathbf{D}|\theta)$ and $F_0(\theta) = \int_{-\infty}^{\theta} d\theta f(\mathbf{D}|\theta)$ and $P(\mathbf{D}) = F_0(\infty)$. If the distribution of $F_0(\theta_i)$ for $\{\theta_i\}$ is not uniform in $[0, 1]$, one can perturb $f(\mathbf{D}|\theta)$ by moving some density from a region of under sampling to a region of over sampling and, thereby, increase $P(\mathbf{D})$.

For an example, I apply the VW method to the following two models defined by a two-dimensional normal distribution and by a power-law-like distribution with unknown centres and widths in each dimension:

$$\begin{aligned} P_{Gauss}(x, y; \theta_x, \theta_y, \sigma_x, \sigma_y) &= \left(\frac{1}{2\pi\sigma_x\sigma_y} \right) e^{-r^2/2}, \quad (23) \\ P_{Power}(x, y; \theta_x, \theta_y, \sigma_x, \sigma_y, \alpha) &= \left(\frac{1}{2\pi\sigma_x\sigma_y} \right) \frac{\alpha(1+\alpha)}{(1+r)^{2+\alpha}} \end{aligned} \quad (24)$$

where $r^2 \equiv x^2/\sigma_x^2 + y^2/\sigma_y^2$ and $\alpha > 0$. For the examples here, I adopt $\alpha = 1$. I take the Gaussian model, P_{Gauss} , or power-law model, P_{Power} , to be the null hypothesis (denoted 0). For each model, I assume the centres are normally distributed with zero mean and unit variance and that the variance is Weibull distributed (eq. B2) with scale 0.03 and shape of 1. A 1000-point data set is sampled from a two-dimensional normal distribution centred at the origin with a root variance of each dimension of 0.03.

I take the same model with the VW extension with $n = 5$ basis functions to be the alternative hypothesis (denoted 1) and test the support for the two hypotheses using Bayes factors. Although VW recommend the choice $w = 1$ for the hyperprior, I found that this tends to overfit the extended model, not favouring the model when it is correct. Rather I adopt $w = 4$ which suppresses this tendency. I performed four MCMC simulations with and without the VW extension and with both the Gaussian and power-law models. Each used the tempered differential evolution algorithm (see §3.1.4) with 32 chains and $T_{max} = 32$ to obtain two million converged states. Then, the posterior samples were batched into groups of 250,000 states and the marginal likelihood was computed using the algorithms described in §3.2. The 90% credible interval was computed by bootstrap analysis from the batches. The results are summarised in Table 3. One sees from the table that the true model is preferred to the extended VW model, but only mildly. Conversely, the power-law model is strongly disfavoured relative to the extended model for the normal data sample. Also note: the marginal likelihood value for the normally distributed data given the normal model (Row 1 in the table) has almost the same value as the extended VW power-law model (Row 4 in the table). As pointed out by VW, the extended model provides an estimator of the true posterior density. The agreement of these two values suggests that the five basis functions provide sufficient variation to reproduce the true value and that the ten-dimensional numerical evaluation of the marginal likelihood integral using the methods described in §3.2 is sufficiently accurate.

4 CASE STUDIES

4.1 Semi-analytic galaxy formation models: BIE-SAM

Many of the physical processes parametrised in semi-analytical models of galaxy formation remain poorly understood and under specified. This has two critically important consequences for inferring constraints on the physical parameters: 1) prior assumptions about the size of the domain and the shape of the parameter distribution will strongly affect any resulting inference; and 2) a very large parameter space must be fully explored to obtain an accurate inference. Both of these issues are naturally tackled with a Bayesian approach that allows one to constrain the theory with data in a probabilistically rigorous way. In Lu et al. (2011), we presented a semi-analytic model (SAM) of galaxy formation in the framework of Bayesian inference and illustrated its performance on a test problem using the BIE; we call the combined approach BIE-SAM. Our sixteen-parameter semi-analytic model incorporates all of the most commonly used parametrizations of important physical processes from existing SAMs including star formation, SN feedback, galaxy mergers, and AGN feedback.

To demonstrate the power of this approach, the thirteen of these parameters that can be constrained by the K-band luminosity function were investigated in Lu et al. (2011). We find that the posterior distribution has a very complex structure and topology, indicating that finding the best fit by tweaking model parameters is improbable. As an example, Figure 2 describes isosurfaces of

the posterior distribution in three of thirteen dimensions. The surfaces have a complex geometry and are strongly inhomogeneous in any parameter direction. Moreover, the posterior clearly shows that many model parameters are strongly covariant and, therefore, the inferred value of a particular parameter can be significantly affected by the priors used for the other parameters. As a consequence, one *may not* tune a small subset of model parameters while keeping other parameters fixed and expect a valid result.

Apropos the discussion in §2.4, by using synthetic data to mimic systematic uncertainties in the reduced data, we also have shown that the resulting model parameter inferences can be significantly affected by the use of an incorrect error model. We used a synthetically-generated binned stellar mass function and performed two inferences: one with a realistic covariance and one with no off-diagonal covariance. The contours with the full covariance matrix are more compact, but there are also noticeable changes in the shape and orientation of the posterior distribution. This clearly demonstrates that an accurate analysis of errors, both sampling errors and systematic uncertainties, are crucial for observational data, and conversely, a data-model comparison without an accurate error model is likely to be erroneous.

The method developed here can be straightforwardly applied to other data sets and to multiple data sets simultaneously. In addition, the Bayesian approach explicitly builds on previous results by incorporating the constraints from previous inferences into new data sets; the BIE is designed to do this automatically. For many processes in galaxy formation, competing models have been proposed but not quantitatively compared. Bayes factor analyses (see §3.2) or explicit model comparison techniques such as the reversible jump algorithm (Green 1995, see also §3.3) can provide a quantitative comparison of different models for a given data set.

4.2 Galphat

Yoon et al. (2011) describes Galphat (GALaxy PHotometric ATtributes), a Bayesian galaxy image analysis package built for the BIE, designed to efficiently and reliably generate the posterior probability distribution of model parameters given an image. From the BIE point of view, both Galphat and the BIE-SAM are likelihood functions. The Galphat likelihood function is designed to produce high-accuracy photometric predictions for galaxy models in an arbitrary spatial orientation for a given point-spread function (PSF). Accurate predictions in both the core and wings of the image are essential for reliable inferences. The pixel predictions are computed using an adaptive quadrature algorithm to achieve a pre-defined error tolerance. The rotation and PSF convolution are performed on sub-sampled grids using an FFT algorithm. Galphat can incorporate any desired galaxy image family. For speed, we pre-compute subsampled grids for each model indexed by parameters that influence their shape. The desired amplitude and linear scale are easily computed through coordinate transformations. To enable this, the images are stored as two-dimensional cumulative distributions. Our current implementation uses Sérsic (1963) models for disk, bulge and spheroid components, however, this approach is directly applicable to any model family.

Using the various tempering algorithms available in the BIE, our tests have demonstrated that we can achieve a steady-state distribution and that the simulated posterior will include any multiple modes consistent with the prior distribution. Given the posterior distribution, we may then consistently estimate the credible regions for the model parameters. We show that the surface-brightness model will often have correlated parameters and, there-

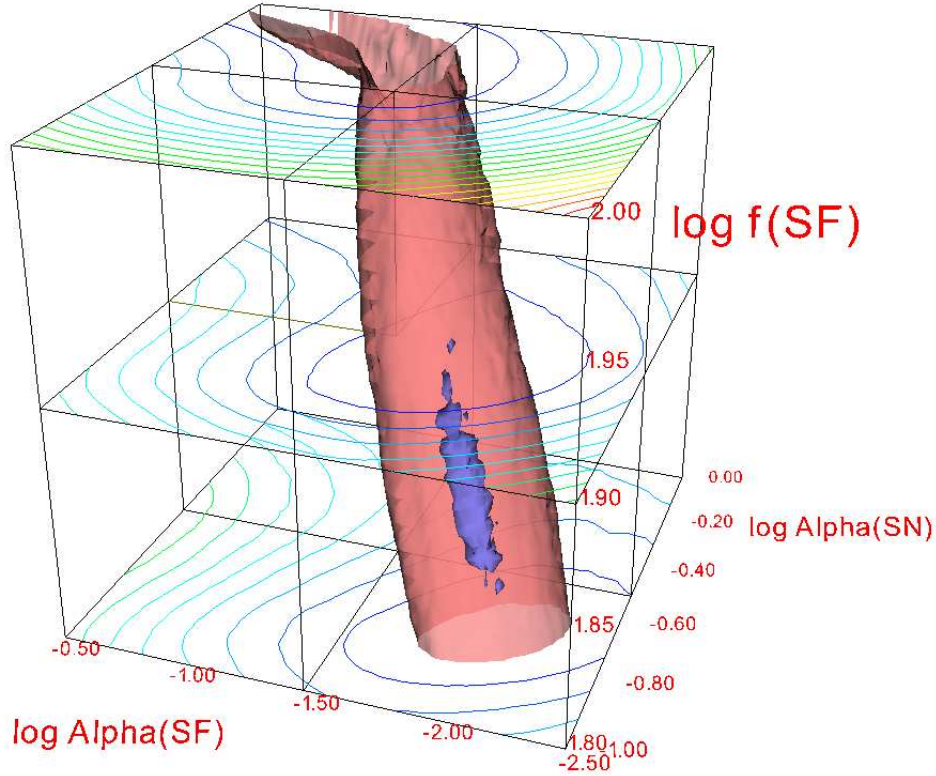


Figure 2. The likelihood function for 3 out of the 13 free parameters in the BIE-SAM from Lu et al. (2011): the star-formation threshold surface density f_{SF} , the star-formation efficiency power-law index α_{SF} , and the supernova feedback energy fraction α_{SN} . The blue (red) surfaces enclose approximately 10% (67%) of the density. See Lu et al. (2011) for additional details.

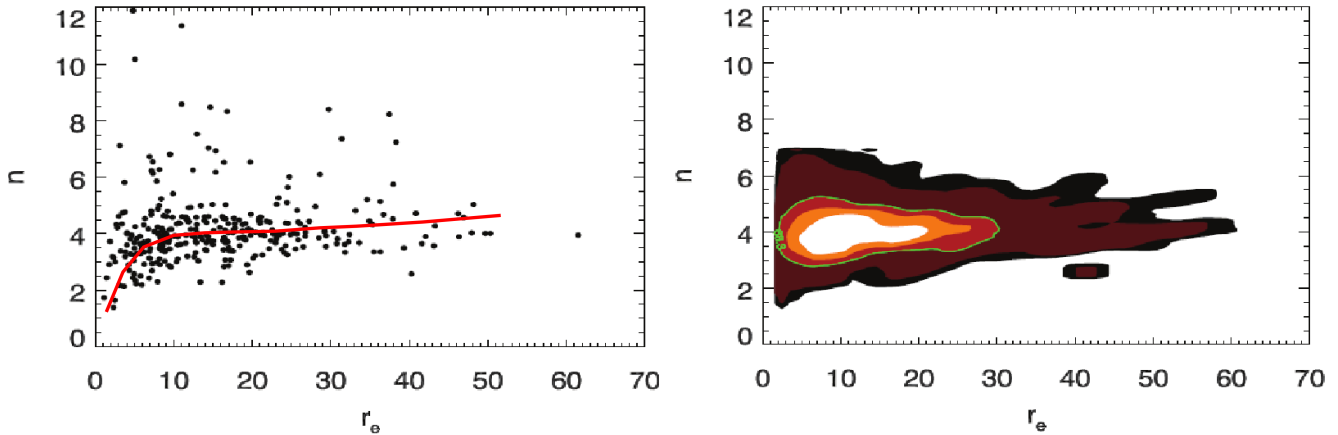


Figure 3. The size–Sérsic index relation inferred from a synthetically-generated sample of elliptical galaxy images. Left: a scatter plot using the best-fit parameters. The red curve shows a smooth fit to the ridge-line of the points. Right: the marginal posterior density for the same parameters. While the left-hand plot suggests that small galaxies have low concentrations, the right-hand plot of posterior density correctly reveals that this trend is an artifact of the model-fitting procedure. These figures originally appeared in Yoon et al. (2011).

fore, any hypothesis testing that uses the ensemble of posterior information will be affected by these correlations. The full posterior distributions from Galphat identify these correlations and incorporate them in subsequent inferences.

These issues are illustrated in Figures 3 and 4. Figure 3 shows the size–Sérsic index relation inferred from a synthetically-generated sample of elliptical galaxy images. The left-hand panel shows the traditional scatter diagram of maximum posterior param-

eter values; that is, this figure plots the effective radius and Sérsic index for the maximum likelihood value obtained for each image fit. The red curve is a smooth estimate of the trend. The right-hand panel shows the inferred distribution based on the full posterior distribution of the ensemble after marginalising over all of the parameters but the effective radius and the Sérsic index. The left-hand panel *incorrectly* suggests that smaller galaxies are less concentrated while the right-hand panel *correctly* reveals that the size and

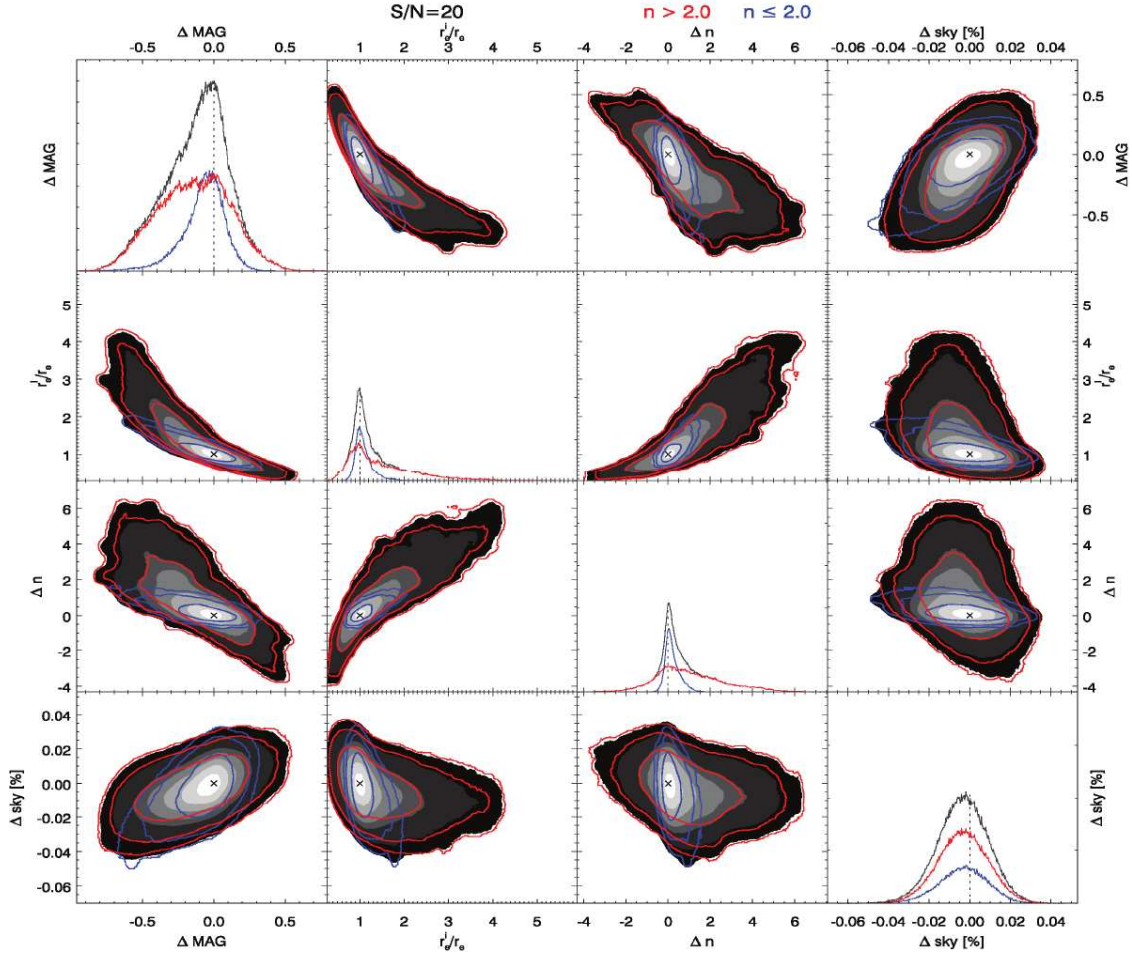


Figure 4. Marginal posterior densities for 100 galaxies with a signal-to-noise ratio of 20 and randomly sampled Sérsic model parameters. The values are magnitude difference from the input values (ΔMAG), galaxy half-light radius scaled by the input values (r_e^i/r_e), the Sérsic index difference (Δn) and the sky value difference (Δsky in percent). The blue, red and grey curves and contours show galaxies with $n \geq 2$, $n < 2$ and the total sample, respectively. The parameter covariance depends on both the signal-to-noise ratio and the Sérsic index.

concentration are uncorrelated. Figure 4 illustrates the correlation between parameters for low-concentration galaxies (Sérsic index $n \leq 2$) in blue and high-concentration galaxies ($n > 2$) in red with the total in grey scale.

We can use posterior simulations over ensembles of images to test, for example, the significance of cluster and field environments on galaxies as evidenced in their photometric parameters, such as the correlation between bulge-to-disk ratio and environment. A more elaborate example might include models for higher angular harmonics of the light distribution and we could determine the support for these in the data using Bayes factors (§3.2).

5 DISCUSSION AND SUMMARY

Advances in digital data acquisition along with storage and retrieval technologies have revolutionised astronomy. The promise of combining these vast archives have led to organisation frameworks such as the National Virtual Observatory (NVO) and the International Virtual Observatory Alliance (IVOA). To realise the promise of this vast distributed repository, the modern astronomer needs and wants to combine multi-sensor data from various surveys to help constrain the complex processes that govern the Universe. The neces-

sary tools are still lacking, and the Bayesian Inference Engine (BIE) was designed as a research tool to fill this gap. This paper outlines the motivation, goals, architecture, and use of the BIE and reports our experience in applying MCMC methods to observational and theoretical Bayesian inference problems in astronomy.

Most researchers are well-versed in the identifying “best” parameters for a particular model for some data using the maximum likelihood method. For example, consider the fit of a surface brightness model to galaxy images. Parameters from the maximum-likelihood solutions are typically plotted in a scatter diagram and trends are interpreted physically. However, plotting scatter diagrams from multiple data sources inadvertently mixes error models and selection effects. §2.1 described the pitfalls of this approach. Rather, the astronomer wants to test the hypothesis that the data is correlated with a coefficient larger than some predetermined value α , a complex *hypothesis test*. However, without incorporating the correlations imposed by both the theoretical model, the error model, and the selection process, the significance of the test is uncertain. Similarly, the astronomer needs methods of assessing whether a posited model is correct. I have divided these needs into two categories: goodness-of-fit tests (§2.2) and model selection (§2.3). As an example of the former, the astronomer may have

found the best parameters using maximum likelihood, but does the model fully explain all of the features in the data? If it does not, one must either modify or reject the model before moving on to the next step. As an example of the latter, suppose an inference results in two parameter regions or multiple models that explain the data. Which model *best* explains the data?

All of these wants and needs—combining data from multiple sources, estimating the probability of model parameters, assessing goodness of fit, and selecting between competing models—are naturally addressed in a single probabilistic framework known as *Bayesian inference*. In particular, Bayesian inference provides a data-first discipline that demands that the error model and selection effects are specified by the probability distribution for the data given the model \mathcal{M} , $P(\mathbf{D}|\theta, \mathcal{M})$, colloquially known as the likelihood function $L(\mathbf{D}|\theta, \mathcal{M})$. Prior results including quantified expert opinion are specified in the prior probability function $\pi(\theta|\mathcal{M})$. The inferential computation may be incremental: the data may be added in steps and new or additional observations may be motivated at each step, true to the scientific method. In the end, this approach may be generalised to locating the most likely models in the generalised space of models; this leads to goodness-of-fit and model comparison tests.

For scientists, the ideal statistical inference is one that lets the data “speak for themselves.” This is achievable in some cases. For example, estimating a small number of parameters given a large data set tends to be independent of prior assumptions. On the other hand, hypothesis tests of two complex models may depend sensitively on prior assumptions. For a trivial example, some choices of parameters may be unphysical even though they yield good fits and should be excluded from consideration. Moreover, if two competing models fit the data equally well, any hypothesis test will be dominated by prior information. Inferences based on realistic models of astronomical systems will often lie between these two extremes. For these cases, I advocate Bayesian methods because they precisely quantify both the scientists’ prior knowledge and the information gained through observation. In other words, we allow the data to “speak for themselves” but in a “dialect” of our choosing. Philosophy aside, Bayes theorem simply embodies the law of conditional probability and provides a rigorous framework for combining the prior and derived information.

With these advantages comes a major disadvantage: Bayesian inference is computationally expensive! An inference may require a huge sample from the posterior distribution and real-world computation of $P(\mathbf{D}|\theta, \mathcal{M})$ is often costly. Moreover, naive MCMC algorithms used for sampling the posterior distribution converge unacceptably slowly for distributions with multiple modes, and advanced techniques to improve the mixing between modes are needed, increasing the expense. Finally, Bayesian inference requires integrals over typically high-dimensional parameter spaces. For example, Bayes factors (§2.3) require the computation of the marginal likelihood:

$$P(\mathbf{D}|\mathcal{M}) = \int d\theta \pi(\theta|\mathcal{M}) P(\mathbf{D}|\theta, \mathcal{M}).$$

Evaluation of this integral suffers from the curse of dimensionality.

Nonetheless, the elegance and promise of Bayesian inference motivated us to attempt a computational solution and this became the BIE project. The algorithms and techniques described here, all and more available in the BIE, have proved useful to address the complications found in research problems. In short, the BIE fills a gap between tools developed for small-scale problems or those designed to test new algorithms and a computational platform designed for production-scale inference problems typical of present-

day astronomical survey science. Its primary product is a sample from a posterior distribution to be used for parameter estimation and model selection. Other Bayesian applications, such as non-parametric inference and clustering, should be possible with little modification, but have not yet been investigated. The BIE is designed to run on high-performance computing clusters, although it will also run on workstations and laptops.

The open object-oriented architecture allows for cross-fertilisation between researchers and groups with both mathematical and scientific interests, e.g. both those developing new algorithms and those developing new astronomical models for different applications. New classes contributed by one become available to all users after an upgrade. Approaches implemented by the one user’s new classes may solve an unanticipated set of problems for other users. In this way, the BIE is a distributed collaborative system similar to packages like IDL or modules in Python. I anticipate that users with a variety of technical skill levels will use the BIE. By reusing and modifying supplied examples, a user’s model of a likelihood function can be straightforwardly added to the system without any detailed knowledge of the internals. A user’s new model becomes an internally-documented first-class object within the BIE by following the examples as templates. There is also room for the experienced programmer to improve the low-level parallelism or implement more efficient heuristics for likelihood evaluations. The BIE includes a full persistence subsystem to save the state and data for running a MCMC simulation. This facilitates both checkpointing and recovery as well as later use of inferred posterior distributions in new and unforeseen ways. A future version will implement a built-in database for warehousing results including the origin and history of both the data and computation, along with labels, notes and comments. Altogether, this will constitute an electronic notebook for Bayesian inference.

This paper provides some concrete examples of parameter estimation and model comparison using the BIE. These do not exhaust the full BIE feature set but, rather, provide a quick introduction to illustrate and motivate its use. The BIE provides the astronomer an organisational and computational schema that discriminates between models or hypotheses and suggests the best use of scarce observational resources. In short, if we can make better use of the interdependency of our observations given our hypotheses, then we can generate a far clearer picture of the underlying physical mechanisms. In many ways, the Bayesian approach emulates the empirical process: begin with a scientific belief or expectation, add the observed data and then modify the extent of that belief to generate the next expectation. In effect, as more and more data is added to the model, the accurate predictions become reinforced and the inaccurate ones rejected. This approach relieves the scientist from directly confronting the complex interdependencies within the data since those interdependencies are automatically incorporated into the model. Although our scientific motivations are astronomical, the BIE can be applied to many different complex systems and may even find applications in areas as diverse as biological systems, climate change, and finance.

6 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. 0611948 and 1109354 and by NASA AISR Program through award NNG06GF25G. The initial development of the BIE was previously funded by NASA’s Applied Information and System Research (AISR) Program. I thank Neal Katz

and Eliot Moss for comments on an early draft of this manuscript
Alison Crocker, Neal Katz, Yu Lu, and Michael Petersen for comments on the final draft.

REFERENCES

- Berntsen J., Espelid T. O., Genz A., 1991, *ACM Trans. Math. Soft.*, 17, 437451 3.2
- Gelman A., 2003, *International Statistical Review*, 71, 369 3.5.1
- Gelman A., Carlin J., Stern H., Rubin D., 1995, *Bayesian Data Analysis*. Chapman and Hall 2.2, 2.3
- Gelman A., Rubin D. B., 1992, *Stat. Sci.*, 7, 457 3.4
- Geyer C. J., 1991, in Keramidas ed., *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface Markov Chain Monte Carlo maximum likelihood*. Interface Foundation, pp 156–163 3.1.2
- Giakoumatos S., Vrontos I., Dellaportas P., Politis D., 1999, *J. Comput. Graph. Statistics*, 8, 431 3.4
- Green P. J., 1995, *Biometrika*, 82, 711 3.3, 3.3, 4.1
- Hastings W. K., 1970, *Biometrika*, 57, 97 3.1.1
- Jeffreys H., 1961, *Theory of Probability* (3rd edition). Oxford University Press 2.3
- Kass R. E., Raftery A. E., 1995, *Journal of the American Statistical Association*, 90, 773 2.3, 3.2
- Kirkpatrick S., Gelatt C. D., Vecchi M. P., 1983, *Science*, 220, 671 3.1.2
- Lewis S. M., Raftery A. E., 1997, *J. Am. Stat. Assoc.*, 440, 648 3.2
- Lindley D. V., 1957, *Biometrika*, 44, 187192 2.3
- Liu J. S., 2004, *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics, Springer 3.1.1
- Lu Y., Mo H. J., Katz N., Weinberg M. D., 2012, *MNRAS*, in press 3.5.1
- Lu Y., Mo H. J., Weinberg M. D., Katz N., 2011, *MNRAS*, 416, 1949 2.4, 3.1.4, 4.1, 2
- Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., Teller E., 1953, 21, 1087 3.1.1, 3.1.1
- Neal R. M., 1996, *Statistics and Computing*, 6, 353 3.1.2, 3.1.2
- Newton M. A., Raftery A. E., 1994, *Journal of the Royal Statistical Society, Ser. B*, 56, 3 3.2
- Pearson K., 1900, *Philosophical Magazine*, 50, 157 1
- Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, *Numerical Recipes in C*, second edn. Cambridge University Press, Cambridge 3.4
- Price K., 1997, *Dr. Dobbs Journal*, 264, 18 3.1.4
- Raftery A. E., 1995, in *Markov Chain Monte Carlo in Practice*. Hypothesis testing and model selection with posterior simulation 2.3
- Robert C., Casella G., 2004, *Monte Carlo Statistical Methods*, second edn. Texts in Statistics, Springer 1
- Sérsic J. L., 1963, *Boletín de la Asociación Argentina de Astronomía La Plata Argentina*, 6, 41 4.2
- Storn K., 1999, in Corne D., Dorigo M., Glover F., eds., *New Ideas in Optimization*. McGraw-Hill, London, pp 79–108 3.1.4
- Storn R., Price K., 1997, *Journal of Global Optimization*, 11, 341 3.1.4
- Ter Braak C. J. F., 2006, *Stat. Comput.*, 16, 239 3.1.4
- Trotta R., 2008, *Contemporary Physics*, 49, 71 1, 2.3
- Verdinelli I., Wasserman L., 1998, *Ann. Statist.*, 26, 1215 2.2, 3.5.2, 3.5.2
- Weinberg M. D., 2012, *Bayesian Analysis*, accepted 2.3, 3.2, 3.2, 3.2
- Yoon I., Weinberg M. D., Katz N. S., 2011, *MNRAS*, 414, 1625 4.2, 3
- Yoon I., Weinberg M. D., Katz N. S., 2012, *MNRAS*, to be submitted 3.1.5, 3.2, 3.2

APPENDIX A: BIE: TECHNICAL OVERVIEW

The BIE is a general-purpose system for simulating Bayesian posterior distributions for statistical inference and has been stress tested using high-dimensional models. As described in the previous sections, the inference approach uses the Bayesian framework enabled by Monte Carlo Markov chain (MCMC) techniques. The software is parallel and multi-threaded and should run in any environment that supports POSIX threads and the widely-implemented Message Passing Interface (MPI, see <http://www.mpi-forum.org>). The package is written in C++ and developed on the GNU/Linux platform but it should port to any GNU platform.

BIE at its core is a software library of interoperable components necessary for performing Bayesian computation. The BIE classes are available as both C++ libraries and as a stand-alone system with an integrated command-line interface. The command-line interface is well tested and is favoured by most users so far. A user does not need to be an expert or even an MPI programmer to use the system; the simple user interface is similar to MatLab or Gnuplot. In addition to the engine itself, the BIE package includes a number of stand-alone programs for viewing and analysing output from the BIE, testing the convergence of a simulation, manipulating the simulation output, and computing the marginal likelihood using the algorithms described in §3.2. A future release will include wrappers for Python.

A1 Software architecture

As in GIMP Toolkit (GTK+) and the Visualisation Toolkit (VTK), the BIE uses C++ to facilitate implementing an object-oriented design. Object-oriented programming enforces an intimate relationship between the data and procedures: the procedures are responsible for manipulating the data to reflect its behaviour. The software objects (classes in C++) represent real-world probability distributions, mathematical operators and algorithms, and this presents a natural interface and set of interobject relationships to the user and the developer. Programs that want to manipulate an object have to be concerned only about which messages this object understands, and do not have to worry about how these tasks are achieved nor the internal structure of the object.

Another powerful feature of object-oriented programming is *inheritance* (derived classes in C++). The derived class inherits the properties of its base class and also adds its own data and routines. This structure makes it easy to make minor changes in the data representation or the procedures. Changes inside a class do not affect any other part of a program, since the only public interface that the external world has to a class is through the use of methods. This facilitates adding new features or responding to changing operating environments by introducing a few new objects and modifying some existing ones. These features encourage extensibility through the reuse of commonly used structures and innovation by allowing the user to connect components in new and possibly unforeseen ways. In addition, this facilitates combining concurrently

developed software contributions from scientists interested in specific models or data types and MCMC algorithms.

Motivated by handling large amounts of survey data with the subsequent possible need to investigate the appropriate simulation for a variety of models or hypotheses, the BIE separates the computation into a collection of subsystems:

- (i) Data input and output
- (ii) Data distribution and spatial location
- (iii) Markov chain simulation
- (iv) Likelihood computation
- (v) Model and hypothesis definition

Each of these can be specified independently and easily mixed and matched in our object-oriented architecture. The cooperative development enabled by this architecture is similar to that behind the open-source movement, and this project is a testament to its success in an interdisciplinary scientific collaboration. The BIE use SVN version management (autoconf, automake), GNU coding standards, and DejaGNU regression testing to aid in portability. Moreover, I believe that this same social model will extend to remote collaborative efforts as the BIE project matures.

A2 Persistence system

The researcher needs to be able to stop, restart, and possibly refocus inferential computations for both technical and scientific reasons. The BIE was designed with these scenarios in mind. The BIE's persistence system is built on top of the BOOST (<http://www.boost.org>) serialisation library. The BIE classes inherit from a base serialisation class that provides the key serialisation members and a simple mnemonic scheme to mark persistent data in newly developed classes. The serialisation-based persistence system avoids irrevocably modifying data sets or files. Rather it views the computation as a series of functions, each accepting one or more input data sets and producing one or more new output data sets. The system records and timestamps these computations and the relationships between inputs and outputs in an archive research log, so that one can always go back and determine the origin of data and how it was processed. The most common use of BIE persistence to date is checkpointing and recovery. Checkpointing guards against loss of computation by saving intermediate data to support recovery in the middle of long-running computational steps; and it allows one to “freeze” or “shelve” a computation and pick it up later. It also provides the basic support needed to interrupt a computation, do some reconfiguring, and resume, as when machines need to be added to or removed from a cluster, etc.

A3 Extensibility

BIE is designed to be extensible. The user may define new classes for any aspect of the MCMC simulation, such as MCMC algorithms, convergence tests, prior distributions, data types, and likelihood functions. The code tree includes a `Projects` directory that is automatically compiled into any local build that may contain any locally added functionality. For example, both the BIE-SAM and Galphat were derived from a base class specifically for user-defined likelihood functions.

The source tree is available for download from <http://www.astro.umass.edu/bie>. The package includes Debian and Ubuntu package management scripts so that local `.deb` packages may be built. Users have had success building other modern Linux distributions.

APPENDIX B: DEFINITION OF DISTRIBUTIONS

B1 Normal distribution

The normal (or Gaussian) distribution is the familiar *bell-shaped* density function that results from the central limit theorem:

$$\mathcal{N}(\theta; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} \quad (\text{B1})$$

where parameter μ is the mean and σ^2 is the variance. See any standard reference in probability and statistics for the multivariate generalisation.

B2 Weibull distribution

The Weibull distribution has a scale parameter (λ) and shape parameter κ . It includes the exponential distribution for $\kappa = 1$ but becomes more peaked around λ as κ increases. The probability density function of a Weibull random variable θ is:

$$\mathcal{W}(\theta; \lambda, \kappa) = \begin{cases} \frac{\kappa}{\lambda} \left(\frac{\theta}{\lambda}\right)^{\kappa-1} e^{-(\theta/\lambda)^\kappa} & \theta \geq 0, \\ 0 & \theta < 0. \end{cases} \quad (\text{B2})$$

B3 Beta distribution

The beta distribution is a two-parameter density defined on the unit interval. The probability density function of the beta distribution is:

$$P_B(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (\text{B3})$$

where $\Gamma(\cdot)$ is the gamma function.

B4 Dirichlet distribution

The Dirichlet distribution is the multivariate r -dimensional generalization of the beta distribution with parameters $\alpha_i, i = 1, \dots, r$. Thinking of each dimension as a separate type of event i , this distribution describes the probability of events where each type of event has been previously observed $\alpha_i - 1$ times. The Dirichlet distribution of order $r \geq 2$ with parameters α_i has the probability density function:

$$P_D(x_1, \dots, x_r; \alpha_1, \dots, \alpha_r) = \frac{\Gamma(\sum_{i=1}^r \alpha_i)}{\prod_{i=1}^r \Gamma(\alpha_i)} \prod_{i=1}^r x_i^{\alpha_i-1} \quad (\text{B4})$$

where $x_1, \dots, x_{r-1} > 0$ and $x_r = 1 - \sum_{i=1}^{r-1} x_i > 0$.